Class Project 3- Text Analytics

Introduction to Big Data & Analytics
CSCI 6444

Group 13- Amish Raj, Arham Ashfaq
Submitted to Professor Stephen Kaisler

1. Dataset Description

The dataset is a book called “A Princess of Mars” by Edgar Rice Burroughs. It is text file containing several chapters. For the purpose of this project, we use chapters I-XI.

2. Objective

Our objective, through this project is to analyze the large text file “A Princess of Mars” and perform Text analytics on it. Through various methods of analysis, we aim to gain an understanding of the text, the sentiment behind it, and make useful observations.

3. Loading the Dataset and Creating a Corpus

We begin our analysis by setting the working directory and loading the dataset.

```
> getwd()
[1] "/Users/amishraj/Documents/Study/Big Data Analytics/Class Project 3/R Projects/APrincessOfMars"
> library(tm)
Loading required package: NLP
> book <- readLines("APrincessOfMars.txt")
```

We separate the chapters as we want to create a corpus where each of the eleven chapters (Chapter I-XI) is a document. We identify the indices using the which method as shown below and separate the chapters by the chapter names, e.g., “Chapter I”, and store the values in a variable for each chapter. We can examine the output to verify if we have correctly separated the chapters.

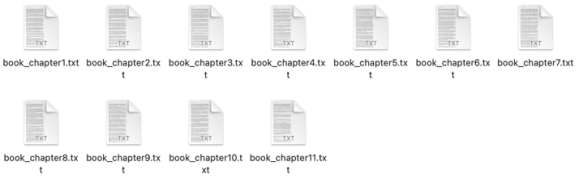
```
> indx_ch1 <- which(book == "CHAPTER I", arr.ind=TRUE)
> indx_ch2 <- which(book == "CHAPTER II", arr.ind=TRUE)
> indx_ch3 <- which(book == "CHAPTER III", arr.ind=TRUE)
> indx_ch4 <- which(book == "CHAPTER IV", arr.ind=TRUE)
> indx_ch5 <- which(book == "CHAPTER V", arr.ind=TRUE)
> indx_ch6 <- which(book == "CHAPTER VI", arr.ind=TRUE)
> indx_ch7 <- which(book == "CHAPTER VII", arr.ind=TRUE)
> indx_ch8 <- which(book == "CHAPTER VIII", arr.ind=TRUE)
> indx_ch9 <- which(book == "CHAPTER IX", arr.ind=TRUE)
> indx_ch10 <- which(book == "CHAPTER X", arr.ind=TRUE)
> indx_ch11 <- which(book == "CHAPTER XI", arr.ind=TRUE)
> indx_ch12 <- which(book== "CHAPTER XII", arr.ind=TRUE)
>
> book_chapter1 <- book[(indx_ch1+1):(indx_ch2-1)]
> book_chapter2 <- book[(indx_ch2+1):(indx_ch3-1)]
> book_chapter3 <- book[(indx_ch3+1):(indx_ch4-1)]
> book_chapter4 <- book[(indx_ch4+1):(indx_ch5-1)]
> book_chapter5 <- book[(indx_ch5+1):(indx_ch6-1)]
> book_chapter6 <- book[(indx_ch6+1):(indx_ch7-1)]
> book_chapter7 <- book[(indx_ch7+1):(indx_ch8-1)]
> book_chapter8 <- book[(indx_ch8+1):(indx_ch9-1)]
> book_chapter9 <- book[(indx_ch9+1):(indx_ch10-1)]
> book_chapter10 <- book[(indx_ch10+1):(indx_ch11-1)]
> book_chapter11 <- book[(indx_ch11+1):(indx_ch12-1)]
>
```

```
> book_chapter1
  [1] ""
  [2] "ON THE ARIZONA HILLS"
  [3] ""
  [4] ""
  [5] "I am a very old man; how old I do not know.  Possibly I am a hundred,"
  [6] "possibly more; but I cannot tell because I have never aged as other"
  [7] "men, nor do I remember any childhood.  So far as I can recollect I have"
  [8] "always been a man, a man of about thirty.  I appear today as I did"
  [9] "forty years and more ago, and yet I feel that I cannot go on living"
 [10] "forever; that some day I shall die the real death from which there is"
 [11] "no resurrection.  I do not know why I should fear death, I who have"
 [12] "died twice and am still alive; but yet I have the same horror of it as"
 [13] "you who have never died, and it is because of this terror of death, I"
 [14] "believe, that I am so convinced of my mortality."
 [15] ""
 [16] "And because of this conviction I have determined to write down the"
 [17] "story of the interesting periods of my life and of my death.  I cannot"
 [18] "explain the phenomena; I can only set down here in the words of an"
 [19] "ordinary soldier of fortune a chronicle of the strange events that"
 [20] "befell me during the ten years that my dead body lay undiscovered in an"
 [21] "Arizona cave."
 [22] ""
 [23] "I have never told this story, nor shall mortal man see this manuscript"
 [24] "until after I have passed over for eternity.  I know that the average"
```

```
> book_chapter11
  [1] ""
  [2] "WITH DEJAH THORIS"
  [3] ""
  [4] ""
  [5] "As we reached the open the two female guards who had been detailed to"
  [6] "watch over Dejah Thoris hurried up and made as though to assume custody"
  [7] "of her once more.  The poor child shrank against me and I felt her two"
  [8] "little hands fold tightly over my arm.  Waving the women away, I"
  [9] "informed them that Sola would attend the captive hereafter, and I"
 [10] "further warned Sarkoja that any more of her cruel attentions bestowed"
 [11] "upon Dejah Thoris would result in Sarkoja's sudden and painful demise."
 [12] ""
 [13] "My threat was unfortunate and resulted in more harm than good to Dejah"
 [14] "Thoris, for, as I learned later, men do not kill women upon Mars, nor"
 [15] "women, men.  So Sarkoja merely gave us an ugly look and departed to"
 [16] "hatch up deviltries against us."
 [17] ""
 [18] "I soon found Sola and explained to her that I wished her to guard Dejah"
 [19] "Thoris as she had guarded me; that I wished her to find other quarters"
 [20] "where they would not be molested by Sarkoja, and I finally informed her"
 [21] "that I myself would take up my quarters among the men."
 [22] ""
 [23] "Sola glanced at the accouterments which were carried in my hand and"
 [24] "slung across my shoulder."
```

We then proceed to create a directory for the chapters and write each chapter to a text file.

```
> dir.create("chapters")
> write.table(book_chapter1, file = "chapters/book_chapter1.txt", sep = "\t", row.names = FALSE, co
l.names = FALSE, quote = FALSE)
> write.table(book_chapter2, file = "chapters/book_chapter2.txt", sep = "\t", row.names = FALSE, co
l.names = FALSE, quote = FALSE)
> write.table(book_chapter3, file = "chapters/book_chapter3.txt", sep = "\t", row.names = FALSE, co
l.names = FALSE, quote = FALSE)
> write.table(book_chapter4, file = "chapters/book_chapter4.txt", sep = "\t", row.names = FALSE, co
l.names = FALSE, quote = FALSE)
> write.table(book_chapter5, file = "chapters/book_chapter5.txt", sep = "\t", row.names = FALSE, co
l.names = FALSE, quote = FALSE)
> write.table(book_chapter6, file = "chapters/book_chapter6.txt", sep = "\t", row.names = FALSE, co
l.names = FALSE, quote = FALSE)
> write.table(book_chapter7, file = "chapters/book_chapter7.txt", sep = "\t", row.names = FALSE, co
l.names = FALSE, quote = FALSE)
> write.table(book_chapter8, file = "chapters/book_chapter8.txt", sep = "\t", row.names = FALSE, co
l.names = FALSE, quote = FALSE)
> write.table(book_chapter9, file = "chapters/book_chapter9.txt", sep = "\t", row.names = FALSE, co
l.names = FALSE, quote = FALSE)
> write.table(book_chapter10, file = "chapters/book_chapter10.txt", sep = "\t", row.names = FALSE,
col.names = FALSE, quote = FALSE)
> write.table(book_chapter11, file = "chapters/book_chapter11.txt", sep = "\t", row.names = FALSE,
col.names = FALSE, quote = FALSE)
```



Once our chapters are separated, we can proceed to create a VCorpus. We utlize the VCorpus() method from the “tm” package to create our corpus and store the corpus in “POM”. The structure can then be examined using the str() method.

```
> POM <- VCorpus(DirSource("./chapters", ignore.case = TRUE, mode="text"))
> str(POM)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 11
  ..$ :List of 2
  .. ..$ content: chr [1:267] "" "ON THE ARIZONA HILLS" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:358] "" "CHAMPION AND CHIEF" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:258] "" "WITH DEJAH THORIS" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter11.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:166] "" "THE ESCAPE OF THE DEAD" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter2.txt"
  .. .. ..$ language     : chr "en"
```

```
> POM
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 11
```

Here, we can see that we have a corpus of 11 documents each representing a chapter and where content represents the text they contain, each having about 160-400 characters. We can also look some other parameters such as meta which contains the metadata of the documents (we haven't specified the author, heading, etc. so those fields appear empty).

If we want to extract text, we can do so in the following manner, as per the rubric. The first document contains 13962 characters. Several lines below have been omitted.

```
> ptext<- POM[[1]]
> ptext
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13962
> ptext[1]
$content
  [1] ""
  [2] "ON THE ARIZONA HILLS"
  [3] ""
  [4] ""
  [5] "I am a very old man; how old I do not know.  Possibly I am a hundred,"
  [6] "possibly more; but I cannot tell because I have never aged as other"
  [7] "men, nor do I remember any childhood.  So far as I can recollect I have"
  [8] "always been a man, a man of about thirty.  I appear today as I did"
  [9] "forty years and more ago, and yet I feel that I cannot go on living"
```

The next step is to create initial Document Term Matrices (DTM). Each row represents a document and each column represents a term in the document. This is helpful as we can analyze our data with vector and matrix algebra. Similarly, we also create a Term Document Matrix (TDM) where rows represent terms and columns represent documents. These are shown below,

utilizing the DocumentTermMatrix() and the TermDocumentMatrix() methods from the “tm” package.

Document Term Matrix:

```
> POM_DTM <- DocumentTermMatrix(POM)
> POM_DTM
<<DocumentTermMatrix (documents: 11, terms: 5003)>>
Non-/sparse entries: 9204/45829
Sparsity           : 83%
Maximal term length: 19
Weighting          : term frequency (tf)
> inspect(POM_DTM)
<<DocumentTermMatrix (documents: 11, terms: 5003)>>
Non-/sparse entries: 9204/45829
Sparsity           : 83%
Maximal term length: 19
Weighting          : term frequency (tf)
Sample             :
                 Terms
Docs                and but for had that the upon was which with
  book_chapter1.txt  90  13  20  25   50 190   12  36    20   25
  book_chapter10.txt 113 18  34  41   50 193   18  54    15   26
  book_chapter11.txt  82  14   9  29   37 119   13  18    13   23
  book_chapter2.txt   58  13  17  20   19 137   13  31    12   10
  book_chapter3.txt   92  15  18  19   26 166   20  36    34   20
  book_chapter4.txt   68  21  10  22   19 148   10  26    16   20
  book_chapter5.txt   56  14  13  14   11  96    8  22     9    8
  book_chapter6.txt   53  10  14  28   11 117   13  17    11   24
  book_chapter7.txt   64   9  18  16   12 166    7  17    22   12
  book_chapter8.txt   81   7  15  24   11 184   22  32    17   11
```

```
> str(POM_DTM)
List of 6
 $ i       : int [1:9204] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:9204] 36 37 38 41 42 60 68 74 78 81 ...
 $ v       : num [1:9204] 1 1 1 1 7 1 1 4 1 1 ...
 $ nrow    : int 11
 $ ncol    : int 5003
 $ dimnames:List of 2
  ..$ Docs : chr [1:11] "book_chapter1.txt" "book_chapter10.txt" "book_chapter11.txt" "book_chapter2.txt" ...
  ..$ Terms: chr [1:5003] "'gentleman'" "\"and" "\"as" "\"because," ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

Examining the DTM, we can notice the 11 rows each representing a chapter, and 5003 terms in total. The maximal term length is 19 and the Sample field gives us a glimpse of the Document Term Matrix. We notice a high sparsity of 83% meaning that there are more terms with 0s in the matrix. This is expected as the corpus is big and not every term would be found in each document.

Term Document Matrix:

```
> POM_TDM <- TermDocumentMatrix(POM)
> POM_TDM
<<TermDocumentMatrix (terms: 5003, documents: 11)>>
Non-/sparse entries: 9204/45829
Sparsity           : 83%
Maximal term length: 19
Weighting          : term frequency (tf)
>
> inspect(POM_TDM)
<<TermDocumentMatrix (terms: 5003, documents: 11)>>
Non-/sparse entries: 9204/45829
Sparsity           : 83%
Maximal term length: 19
Weighting          : term frequency (tf)
Sample             :
       Docs
Terms   book_chapter1.txt book_chapter10.txt book_chapter11.txt book_chapter2.txt book_chapter3.txt book_chapter4.txt
  and                90              113               82              58              92              68
  but                13               18               14              13              15              21
  for                20               34                9              17              18              10
  had                25               41               29              20              19              22
  that               50               50               37              19              26              19
  the               190              193              119             137             166             148
  upon               12               18               13              13              20              10
  was                36               54               18              31              36              26
  which              20               15               13              12              34              16
  with               25               26               23              10              20              20
       Docs
Terms   book_chapter5.txt book_chapter6.txt book_chapter7.txt book_chapter8.txt
  and                56               53               64               81
  but                14               10                9                7
  for                13               14               18               15
  had                14               28               16               24
  that               11               11               12               11
  the                96              117              166              184
  upon                8               13                7               22
  was                22               17               17               32
  which               9               11               22               17
  with                8               24               12               11
```

```
> str(POM_TDM)
List of 6
 $ i       : int [1:9204] 36 37 38 41 42 60 68 74 78 81 ...
 $ j       : int [1:9204] 1 1 1 1 1 1 1 1 1 1 ...
 $ v       : num [1:9204] 1 1 1 1 7 1 1 4 1 1 ...
 $ nrow    : int 5003
 $ ncol    : int 11
 $ dimnames:List of 2
  ..$ Terms: chr [1:5003] "'gentleman'" "\"and" "\"as" "\"because," ...
  ..$ Docs : chr [1:11] "book_chapter1.txt" "book_chapter10.txt" "book_chapter11.txt" "book_chapter2.txt" ...
 - attr(*, "class")= chr [1:2] "TermDocumentMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

The Term Document Matrix gives us similar information in a different matrix.

4. Finding the 10 longest words and sentences in each chapter

Before we cleanse our corpus and remove stopwords, punctuation, etc. we find the 10 longest words and 10 longest sentences in each chapter. We also found the 10 longest words and sentences in all the chapters.

To do this, we utilize the tibbles package that helps us create data frames. To do this, we need to convert the corpus we have into a tibbles data structure. We start by installing the packages "dplyr", "tidytext", and "ggplot2" and making them libraries in our project.

```
> install.packages("dplyr")
also installing the dependency 'pillar'

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/pillar_1.9.0.tgz'
Content type 'application/x-gzip' length 645161 bytes (630 KB)
==================================================
downloaded 630 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/dplyr_1.1.2.tgz'
Content type 'application/x-gzip' length 1590090 bytes (1.5 MB)
==================================================
downloaded 1.5 MB


The downloaded binary packages are in
        /var/folders/d5/w02vn_rj1hv2m765h0gj7qdc0000gn/T//RtmpHuXVVD/downloaded_packages
> install.packages("tidytext")
also installing the dependencies 'janeaustenr', 'tokenizers'

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/janeaustenr_1.0.0.tgz'
Content type 'application/x-gzip' length 1623358 bytes (1.5 MB)
==================================================
downloaded 1.5 MB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/tokenizers_0.3.0.tgz'
Content type 'application/x-gzip' length 952596 bytes (930 KB)
==================================================
downloaded 930 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/tidytext_0.4.1.tgz'
Content type 'application/x-gzip' length 3033144 bytes (2.9 MB)
==================================================
downloaded 2.9 MB


The downloaded binary packages are in
        /var/folders/d5/w02vn_rj1hv2m765h0gj7qdc0000gn/T//RtmpHuXVVD/downloaded_packages
> install.packages("ggplot2")
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/ggplot2_3.4.2.tgz'
Content type 'application/x-gzip' length 4299292 bytes (4.1 MB)
==================================================
downloaded 4.1 MB


The downloaded binary packages are in
        /var/folders/d5/w02vn_rj1hv2m765h0gj7qdc0000gn/T//RtmpHuXVVD/downloaded_packages
```

```
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

> library(tidytext)
> library(ggplot2)

Attaching package: 'ggplot2'

The following object is masked from 'package:NLP':

    annotate
```

Once our libraries have been imported, we convert our corpus "POM" into a tidy Corpus as follows.

```
> tidyPOM <- tidy(POM)
> tidyPOM
# A tibble: 11 × 8
   author datetimestamp       description heading id                language origin text
   <lgl>  <dttm>              <lgl>       <lgl>   <chr>             <chr>    <lgl>  <chr>
 1 NA     2023-05-01 02:13:27 NA          NA      book_chapter1.txt  en       NA     "\nON THE ARIZONA HILLS\n\n\nI am …
 2 NA     2023-05-01 02:13:27 NA          NA      book_chapter10.txt en       NA     "\nCHAMPION AND CHIEF\n\n\nEarly t…
 3 NA     2023-05-01 02:13:27 NA          NA      book_chapter11.txt en       NA     "\nWITH DEJAH THORIS\n\n\nAs we re…
 4 NA     2023-05-01 02:13:27 NA          NA      book_chapter2.txt  en       NA     "\nTHE ESCAPE OF THE DEAD\n\n\nA s…
 5 NA     2023-05-01 02:13:27 NA          NA      book_chapter3.txt  en       NA     "\nMY ADVENT ON MARS\n\n\nI opened…
 6 NA     2023-05-01 02:13:27 NA          NA      book_chapter4.txt  en       NA     "\nA PRISONER\n\n\nWe had gone per…
 7 NA     2023-05-01 02:13:27 NA          NA      book_chapter5.txt  en       NA     "\nI ELUDE MY WATCH DOG\n\n\nSola …
 8 NA     2023-05-01 02:13:27 NA          NA      book_chapter6.txt  en       NA     "\nA FIGHT THAT WON FRIENDS\n\n\nT…
 9 NA     2023-05-01 02:13:27 NA          NA      book_chapter7.txt  en       NA     "\nCHILD-RAISING ON MARS\n\n\nAfte…
10 NA     2023-05-01 02:13:27 NA          NA      book_chapter8.txt  en       NA     "\nA FAIR CAPTIVE FROM THE SKY\n\n…
11 NA     2023-05-01 02:13:27 NA          NA      book_chapter9.txt  en       NA     "\nI LEARN THE LANGUAGE\n\n\nAs I …
```

We can see our tidy corpus contains the metadata we examined earlier along with the text in each document. This is useful in performing our analysis of the 10 longest words and sentences.

First, we compute the 10 longest words and sentences in all the chapters. We do this by tokenizing the words and producing a tibble as follows. The words are stored in POMWords.

```
> POMWords <- tidyPOM %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arrange(desc
(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWords
# A tibble: 23,576 × 3
   id                word            word_length
   <chr>             <chr>                 <int>
 1 book_chapter10.txt responsibilities          16
 2 book_chapter9.txt  responsibilities          16
 3 book_chapter10.txt characteristics           15
 4 book_chapter3.txt  characteristics           15
 5 book_chapter5.txt  characteristics           15
 6 book_chapter1.txt  subconsciously            14
 7 book_chapter1.txt  characteristic            14
 8 book_chapter3.txt  irregularities            14
 9 book_chapter3.txt  characteristic            14
10 book_chapter5.txt  representation            14
# ℹ 23,566 more rows
# ℹ Use `print(n = ...)` to see more rows
>
```

We can see the 10 longest words in all the chapters, highest being the word “responsibilities” with a length of 16 characters, and the 10th word “representation” has 14 characters. Similarly, by modifying the unnest_tokens() method, we obtain the 10 longest sentences in all the chapters.

```
> POMSentences <- tidyPOM %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select(id, se
ntence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentences
# A tibble: 734 × 3
   id                sentence                                                                  sentence_length
   <chr>             <chr>                                                                              <int>
 1 book_chapter11.txt "\nduring the ages of hardships and incessant warring between their own\nvarious …            521
 2 book_chapter11.txt "\n\n\"because, john carter,\" she replied, \"nearly every planet and star\nhavin…            492
 3 book_chapter2.txt  "\n\nfew western wonders are more inspiring than the beauties of an arizona\nmoon…            456
 4 book_chapter8.txt  "  for example,\na proportion of them, always the best marksmen, direct their fir…            454
 5 book_chapter8.txt  "\n\ninstantly the scene changed as by magic; the foremost vessel swung\nbroadsid…            450
 6 book_chapter4.txt  "\n\ni saw no signs of extreme age among them, nor is there any appreciable\ndiff…            449
 7 book_chapter7.txt  "  between these walls the\nlittle martians scampered, wild as deer; being permit…            432
 8 book_chapter10.txt "\n\nwhat words of moment were to have fallen from his lips were never\nspoken, a…            431
 9 book_chapter6.txt  "  my beast had an advantage\nin his first hold, having sunk his mighty fangs far…            431
10 book_chapter10.txt "  numerous brilliantly\ncolored and strangely formed wild flowers dotted the rav…            416
# ℹ 724 more rows
# ℹ Use `print(n = ...)` to see more rows
```

The regular expression above uses ‘.’as a delimiter for identifying the longest sentences, assuming the period symbol is not preceded by any ‘r’ ending titles like, Dr., Mr. etc. This regular expression helps correctly split the text into sentences.

We can see the longest sentence in the chapter 11 having a length of 521 characters and the 10th sentence from chapter 10 has 416 characters.

Next we compute the 10 longest words and sentences for all the chapters. We follow the same procedure except we generate a tidy corpus for each chapter using the tidy() method and then compute the 10 longest words and sentences. These results have been compiled into a table as

follows, and the screenshots of the same have been attached in the following pages for reference. The screenshots display the length of the longest words and sentences found.

10 Longest Words in each chapter:

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7	Chapter 8	Chapter 9	Chapter 10	Chapter 11
1	Subconsciously	Contemplation	Characteristics	Circumstances	Characteristics	overwhelmingly	Representative	Reinforcements	Responsibilities	Responsibilities	Accouterments
2	Characteristic	Metamorphosis	Irregularities	Consideration	Representation	Accomplishing	Conversations	Simultaneously	Expressionless	Characteristics	Circumstances
3	Understanding	Particularly	Characteristic	Manifestation	Ministrations	Gesticulated	Unnecessarily	circumstances	Unintelligible	Companionship	Questioningly
4	comparatively	Predicaments	Consciousness	Consideration	Uncomfortable	Intermediary	Intentionally	Southeasterly	Administration	Manifestation	Eavesdropping
5	sensitiveness	Interruption	Independently	Therapeutics	Intellectual	Transcending	Circumstances	Requisitioned	Accouterments	Ludicrousness	Comparatively
6	Consternation	Overstrained	Accouterments	Scintillated	Intelligence	Commencement	Intellectual	Southwesterly	Importunities	Precipitately	Intermarrying
7	Perpendicular	Bewilderment	Noiselessness	Introduction	Straightaway	Executioner	Conversation	Unaccountably	Possibilities	Authoritative	Irretrievably
8	Resuscitation	Surroundings	gesticulating	Instructions	considerable	Momentarily	Satisfaction	hallucination	Satisfactory	Understanding	Architecture
9	Comparatively	Unfathomable	Comparatively	Observation	Monstrosity	Perceptibly	Appropriated	Majestically	Manufactured	Theoretically	Compositions
10	resurrection	crystallized	interminable	immediately	voluntarily	momentarily	humanitarian	irresistible	intelligence	companionship	conversation

10 Longest Sentences in each chapter:

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7	Chapter 8	Chapter 9	Chapter 10	Chapter 11
1	however, i\nam not prone to sensitiveness, and the following of a sense of duty	\n\nfew western wonders are more inspiring than the beauties of an arizona\nmoonl	the throwing down of his weapons and the\nwithdrawing of his troop before his a…	\n\ni saw no signs of extreme age among them, nor is there any appreciable\ndiffe…	the nights are either brilliantly illumined or\nvery dark, for if neither of th…	my beast had an advantage\nin his first hold, having sunk his mighty fangs far …	between these walls the\nlittle martians scampered, wild as deer; being permitt…	for example,\na proportion of them, always the best marksmen, direct their fire…	oh, it is one continual, awful period of bloodshed from\nthe time we break the …	\n\nwhat words of moment were to have fallen from his lips were never\nspoken, a…	\nduring the ages of hardships and incessant warring between their own\nvarious …
2	the fact that it\nis difficult to aim anything but imprecations accurately by m…	i reasoned with\nmyself that i had lain helpless for many hours within the cave…	he sat his mount as we\nsit a horse, grasping the animal's barrel with his lowe…	they first repeated\nthe word \"sak\" a number of times, and then tars tarkas m…	\n\nthis last device produces an intensely brilliant far-reaching white\nlight, b…	\n\ni am ever willing to stand and fight when the odds are not too\noverwhelming…	\nchild-raising on mars\n\n\nafter a breakfast, which was an exact replica of the…	\n\ninstantly the scene changed as by magic; the foremost vessel swung\nbroadside…	with this added incentive i nearly drove sola distracted by\nmy importunities t…	numerous brilliantly\ncolored and strangely formed wild flowers dotted the rav…	\n\n\"because, john carter,\" she replied, \"nearly every planet and star\nhavin…
3	\n\nsince we had entered the territory we had not seen a hostile indian,\nand we …	\n\nto be held paralyzed, with one's back toward some horrible and unknown\ndange…	their eyes were set at the extreme sides of their heads a\ntriffle above the cen…	\n\nwhat struck me as most remarkable about this assemblage and the hall in\nwhic…	i could not but\nwonder what this ferocious-looking monstrosity might do when I…	\n\ni had at least two friends on mars; a young woman who watched over me\nwith m…	\n\ni do not mean that the adult martians are unnecessarily or\nintentionally cru…	\n\nas sola and i entered the plaza a sight met my eyes which filled my\nwhole be…	\n\n\"when,\" asked one of the women, \"will we enjoy the death throes of the\nre…	i saw\nthat the body of my dead antagonist had been stripped, and i read in\nt…	do not tell me that you have thus returned! they would\nkill you horribly any…
4	\n\nin this instance i was, of course, positive that powell was the center\nof at…	fear is a relative term and so\ni can only measure my feelings at that time by …	\n\nand his mount! how can earthly words describe it! it towered ten feet\nat t…	\n\nmy exhibition had been witnessed by several hundred lesser martians,\nand the…	and it is well that\nnature has so graciously and abundantly lighted the martia…	\n\nsuddenly i came to myself and, with that strange instinct which seems\never t…	entirely\nunknown to their mothers, who, in turn, would have difficulty in\npoi…	\n\nsola and i had entered a building upon the front of the city, in fact,\nthe s…	customs have\nbeen handed down by ages of repetition, but the punishment for ig…	what strange manner\nof man are you, that you consort with the green men, thou	\n\na similar wave of feeling seemed to stir her; she drew away from me\nwith a …
5	\n\nthe morning of powell's departure was, like nearly all arizona\nmornings, cle…	\n\nlate in the afternoon my horse, which had been standing with dragging\nrein b…	\n\nthe respite my unexpected agility had given me permitted me to\nformulate pla…	\n\nthe room was well lighted by a number of large windows and was\nbeautifully d…	it came, as i later discovered, not from\nan animal, as there is only one mamma…	\n\nit is true i held the cudgel, but what could i do with it against his\nfour g…	they were\nnot wanted, as their offspring might inherit and transmit the tenden…	i could not fathom the seeming hallucination, nor could i\nfree myself from it;…	\n\nthe training of myself and the young martians was conducted solely by\nthe wo…	\"\n\n\"then you too are a prisoner? but why, then, those arms and the\nregalia…	i can readily perceive that you are not of\nthe barsoom of today; you are like…
6	\n\nmy horse was traveling practically unguided as i knew that i had\nprobably le…	my\nfirst thought was, is this then death! have i indeed passed over\nforever …	\nthe result is that they are infinitely less agile and less powerful, in\npropor…	owing to\nthe waning resources of the planet it evidently became necessary to\n…	the work had evidently been wrought by a master hand,\nso subtle the atmosphere…	evidently\ndevoid of all the finer sentiments of friendship, love, or affection…	\n\nevery one but myself--men, women, and children--were heavily armed, and\nat t…	whether they had\ndiscovered us or simply were looking at the deserted city i c…	i could not but note the\nunnecessary harshness and brutality with which her gu…	realizing that i was a somewhat favored character, and also\nconvinced that th…	\n\nthe shores of the ancient seas were dotted with just such cities, and\nlesse…
7	\n\ni do not believe that i am made of the stuff which constitutes heroes,\nbecau…	\n\nfrom then until possibly midnight all was silence, the silence of the\ndead; …	\n\ncoming, as they did, over the soft and soundless moss, which covers\npractica…	\nevidently, then, there were other denizens on mars than the wild and\ngrotesque…	\n\nboth of mars' moons are vastly nearer her than is our moon to earth;\nthe nea…	\n\ni was standing near the window and i knew that once in the street i\nmight ga…	\" i saw that he wanted me to repeat my performance of\nyesterday for the edific…	this operation required several\nhours, during which time a number of the chari…	after they had retired for\nthe night it was customary for the adults to carry …	i\nwas soon successful as her injuries amounted to little more than an\nordina…	\n\n\"and whereto, then, would your prisoner escape should you leave her,\nunles…
8	\n\ni was positive now that the trailers were apaches and that they wished\nto ca…	my only alternative seemed to lie\nin flight and my decision was crystallized b…	but the little sound caused me to\nturn, and there upon me, not ten feet from m…	\n\nas he banged me down upon my feet his face was bent close to mine and i\ndid …	\nthe nearer moon of mars makes a complete revolution around the planet\nin a lit…	with a shriek of fear the ape\nwhich held me leaped through the open window, bu…	it is the universal\nlanguage of mars, through the medium of which the higher a…	the sight was\nawe-inspiring in the extreme as one contemplated this mighty flo…	\ni knew that she was fond of me, and now that i had discovered that she\nhated c…	\n\nthe reason for the whole attitude displayed toward me was now apparent;\ni h…	\n\nthese three great divisions of the higher martians had been forced into\na m…
9	\n\ni soon became so drowsy that i could scarcely resist the strong desire\nto th…	there also\ncame to my nostrils a faintly pungent odor, and i could only assume…	\n\ninstead of progressing in a sane and dignified manner, my attempts to\nwalk r…	toward the center\nof the city was a large plaza, and upon this and in the buil…	across the threshold lay stretched the\nsleepless guardian brute, just as i had…	i glimpsed him just before he reached the doorway and the\ninsight of him, now ro…	as i later learned, they had been to the subterranean\nvaults in which the eggs…	\n\nas the craft neared the building, and just before she struck, the\nmartian wa…	they live at peace with all their\nfellows, except when duty calls upon them to…	i could not resist the\nludicrousness of the spectacle, and holding my sides i…	\n\nthese ancient martians had been a highly cultivated and literary race,\nbut …
10	i know that the average\nhuman mind will not believe what it cannot grasp, and …	\n\ni had not long to wait before a stealthy sound apprised me of their\nnearness…	\n\nbehind this first charging demon trailed nineteen others, similar in\nall res…	had the men been strangers, and therefore\nunable to exchange names, they would…	this\ngirl alone, among all the green martians with whom i came in contact,\ndi…	they seemed to be deep in argument, and\ninfinally one of them addressed me, but …	\n\nsola's duties were now doubled, as she was compelled to care for the\nyoung m…	it had never\nbeen given me to see such deadly accuracy of aim, and it seemed a…	\n\ni did not see the prisoner again for several days subsequent to our\nfirst en…	\"\n\nordinarily i am not given to long speeches, nor ever before had i\ndescend…	\nonly in the valley dor, where the river iss empties into the lost sea\nof koru

While these observations are difficult to understand in this form, they give us an insight about some common phrases or sentences in our corpus. We can see that certain words like “responsibilities”, “characteristics” and “momentarily” appear more often in the documents.

```
> tidyPOMCh1<- tidy(POM[1])
> tidyPOMCh1
# A tibble: 1 × 8
  author datetimestamp       description heading id                language origin text
  <lgl>  <dttm>              <lgl>       <lgl>   <chr>             <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter1.txt en       NA     "\nON THE ARIZONA HILLS\n\n\nI am a …
> POMWordsCh1 <- tidyPOMCh1 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arrang
e(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh1 <- tidyPOMCh1 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select
(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh1
# A tibble: 2,616 × 3
   id                word           word_length
   <chr>             <chr>                <int>
 1 book_chapter1.txt subconsciously          14
 2 book_chapter1.txt characteristic          14
 3 book_chapter1.txt understanding           13
 4 book_chapter1.txt comparatively           13
 5 book_chapter1.txt sensitiveness           13
 6 book_chapter1.txt consternation           13
 7 book_chapter1.txt perpendicular           13
 8 book_chapter1.txt resuscitation           13
 9 book_chapter1.txt comparatively           13
10 book_chapter1.txt resurrection            12
# ℹ 2,606 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh1
# A tibble: 79 × 3
   id                sentence                                                                  sentence_length
   <chr>             <chr>                                                                              <int>
 1 book_chapter1.txt "  however, i\nam not prone to sensitiveness, and the following of a sense of duty…         377
 2 book_chapter1.txt "  the fact that it\nis difficult to aim anything but imprecations accurately by m…         372
 3 book_chapter1.txt "\n\nsince we had entered the territory we had not seen a hostile indian,\nand we …         365
 4 book_chapter1.txt "\n\nin this instance i was, of course, positive that powell was the center\nof at…         347
 5 book_chapter1.txt "\n\nthe morning of powell's departure was, like nearly all arizona\nmornings, cle…         326
 6 book_chapter1.txt "\n\nmy horse was traveling practically unguided as i knew that i had\nprobably le…         310
 7 book_chapter1.txt "\n\ni do not believe that i am made of the stuff which constitutes heroes,\nbecau…         289
 8 book_chapter1.txt "\n\ni was positive now that the trailers were apaches and that they wished\nto ca…         279
 9 book_chapter1.txt "\n\ni soon became so drowsy that i could scarcely resist the strong desire\nto th…         273
10 book_chapter1.txt "  i know that the average\nhuman mind will not believe what it cannot grasp, and …         269
# ℹ 69 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> tidyPOMCh2<- tidy(POM[4])
> tidyPOMCh2
# A tibble: 1 × 8
  author datetimestamp       description heading id                language origin text
  <lgl>  <dttm>              <lgl>       <lgl>   <chr>             <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter2.txt en       NA     "\nTHE ESCAPE OF THE DEAD\n\n\nA sen…
> POMWordsCh2 <- tidyPOMCh2 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arrang
e(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh2 <- tidyPOMCh2 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select
(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh2
# A tibble: 1,683 × 3
   id                word          word_length
   <chr>             <chr>               <int>
 1 book_chapter2.txt contemplation          13
 2 book_chapter2.txt metamorphosis          13
 3 book_chapter2.txt particularly           12
 4 book_chapter2.txt predicaments           12
 5 book_chapter2.txt interruption           12
 6 book_chapter2.txt overstrained           12
 7 book_chapter2.txt bewilderment           12
 8 book_chapter2.txt surroundings           12
 9 book_chapter2.txt unfathomable           12
10 book_chapter2.txt crystallized           12
# ℹ 1,673 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh2
# A tibble: 52 × 3
   id                sentence                                                                  sentence_length
   <chr>             <chr>                                                                              <int>
 1 book_chapter2.txt "\n\nfew western wonders are more inspiring than the beauties of an arizona\nmoonl…         456
 2 book_chapter2.txt "  i reasoned with\nmyself that i had lain helpless for many hours within the cave…         394
 3 book_chapter2.txt "\n\nto be held paralyzed, with one's back toward some horrible and unknown\ndange…         373
 4 book_chapter2.txt "  fear is a relative term and so\ni can only measure my feelings at that time by …         359
 5 book_chapter2.txt "\n\nlate in the afternoon my horse, which had been standing with dragging\nrein b…         352
 6 book_chapter2.txt "  my\nfirst thought was, is this then death!  have i indeed passed over\nforever …         275
 7 book_chapter2.txt "\n\nfrom then until possibly midnight all was silence, the silence of the\ndead; …         264
 8 book_chapter2.txt "  my only alternative seemed to lie\nin flight and my decision was crystallized b…         244
 9 book_chapter2.txt "  there also\ncame to my nostrils a faintly pungent odor, and i could only assume…         217
10 book_chapter2.txt "\n\ni had not long to wait before a stealthy sound apprised me of their\nnearness…         213
# ℹ 42 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> tidyPOMCh3<- tidy(POM[5])
> tidyPOMCh3
# A tibble: 1 × 8
  author datetimestamp       description heading id                language origin text
  <lgl>  <dttm>              <lgl>       <lgl>   <chr>             <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter3.txt en       NA     "\nMY ADVENT ON MARS\n\n\nI opened m…
> POMWordsCh3 <- tidyPOMCh3 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arrang
e(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh3 <- tidyPOMCh3 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select
(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh3
# A tibble: 2,609 × 3
   id                word            word_length
   <chr>             <chr>                 <int>
 1 book_chapter3.txt characteristics          15
 2 book_chapter3.txt irregularities           14
 3 book_chapter3.txt characteristic           14
 4 book_chapter3.txt consciousness            13
 5 book_chapter3.txt independently            13
 6 book_chapter3.txt accouterments            13
 7 book_chapter3.txt noiselessness            13
 8 book_chapter3.txt gesticulating            13
 9 book_chapter3.txt comparatively            13
10 book_chapter3.txt interminable             12
# ℹ 2,599 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh3
# A tibble: 86 × 3
   id                sentence                                                                  sentence_length
   <chr>             <chr>                                                                              <int>
 1 book_chapter3.txt "  the throwing down of his weapons and the\nwithdrawing of his troop before his a…         408
 2 book_chapter3.txt "  he sat his mount as we\nsit a horse, grasping the animal's barrel with his lowe…         336
 3 book_chapter3.txt "  their eyes were set at the extreme sides of their heads a\ntrifle above the cen…         335
 4 book_chapter3.txt "\n\nand his mount!  how can earthly words describe it!  it towered ten feet\nat t…         305
 5 book_chapter3.txt "\n\nthe respite my unexpected agility had given me permitted me to\nformulate pla…         288
 6 book_chapter3.txt "\nthe result is that they are infinitely less agile and less powerful, in\nproport…         283
 7 book_chapter3.txt "\n\ncoming, as they did, over the soft and soundless moss, which covers\ngractica…         280
 8 book_chapter3.txt "  but the little sound caused me to\nturn, and there upon me, not ten feet from m…         263
 9 book_chapter3.txt "\n\ninstead of progressing in a sane and dignified manner, my attempts to\nwalk r…         254
10 book_chapter3.txt "\n\nbehind this first charging demon trailed nineteen others, similar in\nall res…         251
# ℹ 76 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> tidyPOMCh4<- tidy(POM[6])
> tidyPOMCh4
# A tibble: 1 × 8
  author datetimestamp       description heading id                language origin text
  <lgl>  <dttm>              <lgl>       <lgl>   <chr>             <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter4.txt en       NA     "\nA PRISONER\n\n\nWe had gone perha…
>
> POMWordsCh4 <- tidyPOMCh4 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arrang
e(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh4 <- tidyPOMCh4 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select
(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh4
# A tibble: 2,112 × 3
   id                word           word_length
   <chr>             <chr>                <int>
 1 book_chapter4.txt circumstances          13
 2 book_chapter4.txt consideration          13
 3 book_chapter4.txt manifestation          13
 4 book_chapter4.txt consideration          13
 5 book_chapter4.txt therapeutics           12
 6 book_chapter4.txt scintillated           12
 7 book_chapter4.txt introduction           12
 8 book_chapter4.txt instructions           12
 9 book_chapter4.txt observation            11
10 book_chapter4.txt immediately            11
# ℹ 2,102 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh4
# A tibble: 69 × 3
   id                sentence                                                                  sentence_length
   <chr>             <chr>                                                                              <int>
 1 book_chapter4.txt "\n\ni saw no signs of extreme age among them, nor is there any appreciable\ndiffe…         449
 2 book_chapter4.txt "  they first repeated\nthe word \"sak\" a number of times, and then tars tarkas m…         411
 3 book_chapter4.txt "\n\nwhat struck me as most remarkable about this assemblage and the hall in\nwhic…         410
 4 book_chapter4.txt "\n\nmy exhibition had been witnessed by several hundred lesser martians,\nand the…         371
 5 book_chapter4.txt "\n\nthe room was well lighted by a number of large windows and was\nbeautifully d…         357
 6 book_chapter4.txt "  owing to\nthe waning resources of the planet it evidently became necessary to\n…         353
 7 book_chapter4.txt "\nevidently, then, there were other denizens on mars than the wild and\ngrotesque…         306
 8 book_chapter4.txt "\n\nas he banged me down upon my feet his face was bent close to mine and i\ndid …         285
 9 book_chapter4.txt "  toward the center\nof the city was a large plaza, and upon this and in the buil…         276
10 book_chapter4.txt "  had the men been strangers, and therefore\nunable to exchange names, they would…         271
# ℹ 59 more rows
# ℹ Use `print(n = ...)` to see more rows
```

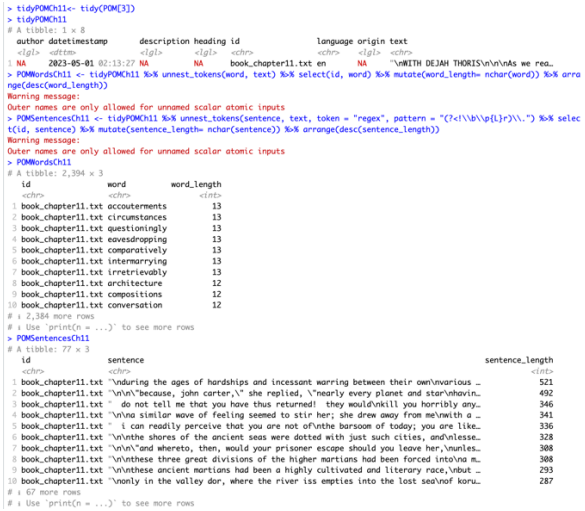
```
> tidyPOMCh5<- tidy(POM[7])
> tidyPOMCh5
# A tibble: 1 × 8
  author datetimestamp       description heading id              language origin text
  <lgl>  <dttm>               <lgl>       <lgl>   <chr>           <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter5.txt en       NA     "\nI ELUDE MY WATCH DOG\n\n\nSola st…
> POMWordsCh5 <- tidyPOMCh5 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arrang
e(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh5 <- tidyPOMCh5 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select
(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh5
# A tibble: 1,538 × 3
   id              word            word_length
   <chr>           <chr>                 <int>
 1 book_chapter5.txt characteristics          15
 2 book_chapter5.txt representation           14
 3 book_chapter5.txt ministrations            13
 4 book_chapter5.txt uncomfortable            13
 5 book_chapter5.txt intellectual             12
 6 book_chapter5.txt intelligence             12
 7 book_chapter5.txt straightaway             12
 8 book_chapter5.txt considerable             12
 9 book_chapter5.txt monstrosity              11
10 book_chapter5.txt voluntarily              11
# ℹ 1,528 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh5
# A tibble: 46 × 3
   id              sentence                                                                          sentence_length
   <chr>           <chr>                                                                                      <int>
 1 book_chapter5.txt "  the nights are either brilliantly illumined or\nvery dark, for if neither of th…            392
 2 book_chapter5.txt "\n\nthis last device produces an intensely brilliant far-reaching white\nlight, b…            367
 3 book_chapter5.txt "  i could not but\nwonder what this ferocious-looking monstrosity might do when l…            348
 4 book_chapter5.txt "  and it is well that\nnature has so graciously and abundantly lighted the martia…            346
 5 book_chapter5.txt "  it came, as i later discovered, not from\nan animal, as there is only one mamma…            312
 6 book_chapter5.txt "  the work had evidently been wrought by a master hand,\nso subtle the atmosphere…            280
 7 book_chapter5.txt "\n\nboth of mars' moons are vastly nearer her than is our moon to earth;\nthe nea…            280
 8 book_chapter5.txt "\nthe nearer moon of mars makes a complete revolution around the planet\nin a lit…            275
 9 book_chapter5.txt "  across the threshold lay stretched the\nsleepless guardian brute, just as i had…            268
10 book_chapter5.txt "  this\ngirl alone, among all the green martians with whom i came in contact,\ndi…            267
# ℹ 36 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> tidyPOMCh6<- tidy(POM[8])
> tidyPOMCh6
# A tibble: 1 × 8
  author datetimestamp       description heading id              language origin text
  <lgl>  <dttm>               <lgl>       <lgl>   <chr>           <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter6.txt en       NA     "\nA FIGHT THAT WON FRIENDS\n\n\nThe…
> POMWordsCh6 <- tidyPOMCh6 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arran
ge(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh6 <- tidyPOMCh6 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select
(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh6
# A tibble: 1,633 × 3
   id              word            word_length
   <chr>           <chr>                 <int>
 1 book_chapter6.txt overwhelmingly           14
 2 book_chapter6.txt accomplishing            13
 3 book_chapter6.txt gesticulated             12
 4 book_chapter6.txt intermediary             12
 5 book_chapter6.txt transcending             12
 6 book_chapter6.txt commencement             12
 7 book_chapter6.txt executioner              11
 8 book_chapter6.txt momentarily              11
 9 book_chapter6.txt perceptibly              11
10 book_chapter6.txt momentarily              11
# ℹ 1,623 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh6
# A tibble: 50 × 3
   id              sentence                                                                          sentence_length
   <chr>           <chr>                                                                                      <int>
 1 book_chapter6.txt "  my beast had an advantage\nin his first hold, having sunk his mighty fangs far …            431
 2 book_chapter6.txt "\n\ni am ever willing to stand and fight when the odds are not too\noverwhelmingl…            378
 3 book_chapter6.txt "\n\ni had at least two friends on mars; a young woman who matched over me\nwith m…            343
 4 book_chapter6.txt "\n\nsuddenly i came to myself and, with that strange instinct which seems\never t…            338
 5 book_chapter6.txt "\n\nit is true i held the cudgel, but what could i do with it against his\nfour g…            297
 6 book_chapter6.txt "  evidently\ndevoid of all the finer sentiments of friendship, love, or affection…            291
 7 book_chapter6.txt "\n\ni was standing near the window and i knew that once in the street i\nmight ga…            265
 8 book_chapter6.txt "  with a shriek of fear the ape\nwhich held me leaped through the open window, bu…            256
 9 book_chapter6.txt "  i glimpsed him just before he reached the doorway and the\nsight of him, now ro…            253
10 book_chapter6.txt "  they seemed to be deep in argument, and\nfinally one of them addressed me, but …            248
# ℹ 40 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> tidyPOMCh7<- tidy(POM[9])
> tidyPOMCh7
# A tibble: 1 × 8
  author datetimestamp       description heading id              language origin text
  <lgl>  <dttm>               <lgl>       <lgl>   <chr>           <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter7.txt en       NA     "\nCHILD-RAISING ON MARS\n\n\nAfter …
> POMWordsCh7 <- tidyPOMCh7 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arran
ge(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh7 <- tidyPOMCh7 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select
(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh7
# A tibble: 2,010 × 3
   id              word            word_length
   <chr>           <chr>                 <int>
 1 book_chapter7.txt representative           14
 2 book_chapter7.txt conversations            13
 3 book_chapter7.txt unnecessarily            13
 4 book_chapter7.txt intentionally            13
 5 book_chapter7.txt circumstances            13
 6 book_chapter7.txt intellectual             12
 7 book_chapter7.txt conversation             12
 8 book_chapter7.txt satisfaction             12
 9 book_chapter7.txt appropriated             12
10 book_chapter7.txt humanitarian             12
# ℹ 2,000 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh7
# A tibble: 62 × 3
   id              sentence                                                                          sentence_length
   <chr>           <chr>                                                                                      <int>
 1 book_chapter7.txt "  between these walls the\nlittle martians scampered, wild as deer; being permitt…            432
 2 book_chapter7.txt "\nchild-raising on mars\n\nafter a breakfast, which was an exact replica of the…            372
 3 book_chapter7.txt "\n\ni do not mean that the adult martians are unnecessarily or\nintentionally cru…            332
 4 book_chapter7.txt "  entirely\nunknown to their mothers, who, in turn, would have difficulty in\npoi…            277
 5 book_chapter7.txt "  they were\nnot wanted, as their offspring might inherit and transmit the tenden…            274
 6 book_chapter7.txt "\n\nevery one but myself--men, women, and children--were heavily armed, and\nat t…            271
 7 book_chapter7.txt "\"  i saw that he wanted me to repeat my performance of\nyesterday for the edific…            270
 8 book_chapter7.txt "  it is the universal\nlanguage of mars, through the medium of which the higher a…            268
 9 book_chapter7.txt "  as i later learned, they had been to the subterranean\nvaults in which the eggs…            264
10 book_chapter7.txt "\n\nsola's duties were now doubled, as she was compelled to care for the\nyoung m…            262
# ℹ 52 more rows
# ℹ Use `print(n = ...)` to see more rows
 > tidyPOMCh9<- tidy(POM[11])
> tidyPOMCh9
# A tibble: 1 × 8
  author datetimestamp       description heading id              language origin text
  <lgl>  <dttm>               <lgl>       <lgl>   <chr>           <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter9.txt en       NA     "\nI LEARN THE LANGUAGE\n\n\nAs I ca…
>
> POMWordsCh9 <- tidyPOMCh9 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arrang
e(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh9 <- tidyPOMCh9 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select
(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh9
# A tibble: 1,429 × 3
   id              word            word_length
   <chr>           <chr>                 <int>
 1 book_chapter9.txt responsibilities         16
 2 book_chapter9.txt expressionless           14
 3 book_chapter9.txt unintelligible           14
 4 book_chapter9.txt administration           14
 5 book_chapter9.txt accouterments            13
 6 book_chapter9.txt importunities            13
 7 book_chapter9.txt possibilities            13
 8 book_chapter9.txt satisfactory             12
 9 book_chapter9.txt manufactured             12
10 book_chapter9.txt intelligence             12
# ℹ 1,419 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh9
# A tibble: 46 × 3
   id              sentence                                                                          sentence_length
   <chr>           <chr>                                                                                      <int>
 1 book_chapter9.txt "  oh, it is one continual, awful period of bloodshed from\nthe time we break the …            315
 2 book_chapter9.txt "  with this added incentive i nearly drove sola distracted by\my importunities t…            285
 3 book_chapter9.txt "\n\n\"when,\" asked one of the women, \"will we enjoy the death throes of the\nre…            281
 4 book_chapter9.txt "  customs have\nbeen handed down by ages of repetition, but the punishment for ig…            280
 5 book_chapter9.txt "\n\nthe training of myself and the young martians was conducted solely by\nthe wo…            274
 6 book_chapter9.txt "  i could not but note the\nunnecessary harshness and brutality with which her gu…            274
 7 book_chapter9.txt "  after they had retired for\nthe night it was customary for the adults to carry …            274
 8 book_chapter9.txt "\ni knew that she was fond of me, and now that i had discovered that she\nhated c…            261
 9 book_chapter9.txt "  they live at peace with all their\nfellows, except when duty calls upon them to…            260
10 book_chapter9.txt "\n\ni did not see the prisoner again for several days subsequent to our\nfirst en…            245
# ℹ 36 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> tidyPOMCh8<- tidy(POM[10])
> tidyPOMCh8
# A tibble: 1 × 8
  author datetimestamp       description heading id              language origin text
  <lgl>  <dttm>               <lgl>       <lgl>   <chr>           <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter8.txt en       NA     "\nA FAIR CAPTIVE FROM THE SKY\n\n\n…
> POMWordsCh8 <- tidyPOMCh8 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arrang
e(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh8 <- tidyPOMCh8 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% select
(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh8
# A tibble: 2,030 × 3
   id              word            word_length
   <chr>           <chr>                 <int>
 1 book_chapter8.txt reinforcements           14
 2 book_chapter8.txt simultaneously           14
 3 book_chapter8.txt circumstances            13
 4 book_chapter8.txt southeasterly            13
 5 book_chapter8.txt requisitioned            13
 6 book_chapter8.txt southwesterly            13
 7 book_chapter8.txt unaccountably            13
 8 book_chapter8.txt hallucination            13
 9 book_chapter8.txt majestically             12
10 book_chapter8.txt irresistible             12
# ℹ 2,020 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh8
# A tibble: 59 × 3
   id              sentence                                                                          sentence_length
   <chr>           <chr>                                                                                      <int>
 1 book_chapter8.txt "  for example,\na proportion of them, always the best marksmen, direct their fire…            454
 2 book_chapter8.txt "\n\ninstantly the scene changed as by magic; the foremost vessel swung\nbroadside…            450
 3 book_chapter8.txt "\n\nas sola and i entered the plaza a sight met my eyes which filled my\nwhole be…            409
 4 book_chapter8.txt "\n\nsola and i had entered a building upon the front of the city, in fact,\nthe s…            346
 5 book_chapter8.txt "  i could not fathom the seeming hallucination, nor could i\nfree myself from it;…            344
 6 book_chapter8.txt "  whether they had\ndiscovered us or simply were looking at the deserted city i c…            343
 7 book_chapter8.txt "  this operation required several\nhours, during which time a number of the chari…            336
 8 book_chapter8.txt "  the sight was\nawe-inspiring in the extreme as one contemplated this mighty flo…            331
 9 book_chapter8.txt "\n\nas the craft neared the building, and just before she struck, the\nmartian wa…            305
10 book_chapter8.txt "  it had never\nbeen given me to see such deadly accuracy of aim, and it seemed a…            295
# ℹ 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> tidyPOMCh10<- tidy(POM[2])
> tidyPOMCh10
# A tibble: 1 × 8
  author datetimestamp       description heading id              language origin text
  <lgl>  <dttm>               <lgl>       <lgl>   <chr>           <chr>    <lgl>  <chr>
1 NA     2023-05-01 02:13:27 NA          NA      book_chapter10.txt en       NA     "\nCHAMPION AND CHIEF\n\n\nEarly th…
> POMWordsCh10 <- tidyPOMCh10 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arra
nge(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh10 <- tidyPOMCh10 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% selec
t(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh10
# A tibble: 3,522 × 3
   id              word            word_length
   <chr>           <chr>                 <int>
 1 book_chapter10.txt responsibilities         16
 2 book_chapter10.txt characteristics          15
 3 book_chapter10.txt companionship            13
 4 book_chapter10.txt manifestation            13
 5 book_chapter10.txt ludicrousness            13
 6 book_chapter10.txt precipitately            13
 7 book_chapter10.txt authoritative            13
 8 book_chapter10.txt understanding            13
 9 book_chapter10.txt theoretically            13
10 book_chapter10.txt companionship            13
# ℹ 3,512 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh10
# A tibble: 108 × 3
   id              sentence                                                                          sentence_length
   <chr>           <chr>                                                                                      <int>
 1 book_chapter10.txt "\n\nwhat words of moment were to have fallen from his lips were never\nspoken, a…            431
 2 book_chapter10.txt "  numerous brilliantly\ncolored and strangely formed wild flowers dotted the rav…            416
 3 book_chapter10.txt "  i saw\nthat the body of my dead antagonist had been stripped, and i read in\nt…            416
 4 book_chapter10.txt "  what strange manner\nof man are you, that you consort with the green men, thou…            405
 5 book_chapter10.txt "\"\n\n\"then you too are a prisoner?  but why, then, those arms and the\nregalia…            393
 6 book_chapter10.txt "  realizing that i was a somewhat favored character, and also\nconvinced that th…            387
 7 book_chapter10.txt "  i\nwas soon successful as her injuries amounted to little more than an\nordina…            353
 8 book_chapter10.txt "\n\nthe reason for the whole attitude displayed toward me was now apparent;\ni h…            345
 9 book_chapter10.txt "  i could not resist the\nludicrousness of the spectacle, and holding my sides i…            336
10 book_chapter10.txt "\"\n\nordinarily i am not given to long speeches, nor ever before had i\ndescend…            334
# ℹ 98 more rows
# ℹ Use `print(n = ...)` to see more rows
```



```
> tidyPOMCh11<- tidy(POM[3])
> tidyPOMCh11
# A tibble: 1 × 8
  author datetimestamp       description heading id               language origin text
  <lgl>  <dttm>              <lgl>       <lgl>   <chr>            <chr>    <lgl>  <chr>
NA     2023-05-01 02:13:27 NA          NA      book_chapter11.txt en       NA     "\nWITH DEJAH THORIS\n\n\nAs we rea…
> POMWordsCh11 <- tidyPOMCh11 %>% unnest_tokens(word, text) %>% select(id, word) %>% mutate(word_length= nchar(word)) %>% arra
nge(desc(word_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMSentencesCh11 <- tidyPOMCh11 %>% unnest_tokens(sentence, text, token = "regex", pattern = "(?<!\\b\\p{L}r)\\.") %>% selec
t(id, sentence) %>% mutate(sentence_length= nchar(sentence)) %>% arrange(desc(sentence_length))
Warning message:
Outer names are only allowed for unnamed scalar atomic inputs
> POMWordsCh11
# A tibble: 2,394 × 3
   id                word          word_length
   <chr>             <chr>               <int>
 1 book_chapter11.txt accouterments          13
 2 book_chapter11.txt circumstances          13
 3 book_chapter11.txt questioningly          13
 4 book_chapter11.txt eavesdropping          13
 5 book_chapter11.txt comparatively          13
 6 book_chapter11.txt intermarrying          13
 7 book_chapter11.txt irretrievably          13
 8 book_chapter11.txt architecture           12
 9 book_chapter11.txt compositions           12
10 book_chapter11.txt conversation           12
# ℹ 2,384 more rows
# ℹ Use `print(n = ...)` to see more rows
> POMSentencesCh11
# A tibble: 77 × 3
   id                sentence                                                                    sentence_length
   <chr>             <chr>                                                                                 <int>
 1 book_chapter11.txt "\nduring the ages of hardships and incessant warring between their own\nvarious …            521
 2 book_chapter11.txt "\n\n\"because, john carter,\" she replied, \"nearly every planet and star\nhavin…            492
 3 book_chapter11.txt "  do not tell me that you have thus returned!  they would\nkill you horribly any…            346
 4 book_chapter11.txt "\n\na similar wave of feeling seemed to stir her; she drew away from me\nwith a …            341
 5 book_chapter11.txt "  i can readily perceive that you are not of\nthe barsoom of today; you are like…            336
 6 book_chapter11.txt "\n\nthe shores of the ancient seas were dotted with just such cities, and\nlesse…            328
 7 book_chapter11.txt "\n\n\"and whereto, then, would your prisoner escape should you leave her,\nunles…            308
 8 book_chapter11.txt "\n\nthese three great divisions of the higher martians had been forced into\na m…            308
 9 book_chapter11.txt "\n\nthese ancient martians had been a highly cultivated and literary race,\nbut …            293
10 book_chapter11.txt "\nonly in the valley dor, where the river iss empties into the lost sea\nof koru…            287
# ℹ 67 more rows
# ℹ Use `print(n = ...)` to see more rows
```

5. Corpus Cleansing/ Data Wrangling

Next, we proceed to clean our data by removing stop words, numbers, and punctuation. Our document is already in lower case. We start by removing quotes from our corpus, and employ gsub() to do so.

```
> POM <- tm_map(POM, content_transformer(gsub), pattern = "'", replacement = "")
```

We then define a function “removeNumPunct()” to remove numbers and punctuation. The POM corpus can then be passed to this function to generate our clean corpus “POMcl” as shown below.

```
> removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*","", x)
> POMcl <- tm::tm_map(POM, content_transformer(removeNumPunct))
> POMcl
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 11
> str(POMcl)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 11
  ..$ :List of 2
  .. ..$ content: chr [1:267] "" "ON THE ARIZONA HILLS" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:358] "" "CHAMPION AND CHIEF" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:258] "" "WITH DEJAH THORIS" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter11.txt"
```

```
> inspect(POMcl)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 11

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13711

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18669

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 12812

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 8844

[[5]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13921

[[6]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 11591

[[7]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 8210

[[8]]
<<PlainTextDocument>>
```

If we inspect our clean corpus, we can notice that we removed 246, 436, 335, 155, 267, 236, 170, 186, 206, 199, and 146 characters from our respective documents representing the chapters. We can make sure our clean corpus is in lowercase by passing it to the tm_map() function.

```
> POMLow<- tm_map(POMcl, tm::content_transformer(tolower))
> POMLow
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 11
> str(POMLow)
Classes 'VCorpus', 'Corpus'  hidden list of 3
 $ content:List of 11
  ..$ :List of 2
  .. ..$ content: chr [1:267] "" "on the arizona hills" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter1.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:358] "" "champion and chief" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter10.txt"
  .. .. ..$ language     : chr "en"
  .. .. ..$ origin       : chr(0)
  .. .. ..- attr(*, "class")= chr "TextDocumentMeta"
  .. ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
  ..$ :List of 2
  .. ..$ content: chr [1:258] "" "with dejah thoris" "" "" ...
  .. ..$ meta   :List of 7
  .. .. ..$ author       : chr(0)
  .. .. ..$ datetimestamp: POSIXlt[1:1], format: "2023-05-01 02:13:27"
  .. .. ..$ description  : chr(0)
  .. .. ..$ heading      : chr(0)
  .. .. ..$ id           : chr "book_chapter11.txt"
  .. .. ..$ language     : chr "en"
```

```
> inspect(POMLow)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 11

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13711

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18669

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 12812

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 8844

[[5]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 13921

[[6]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 11591

[[7]]
<<PlainTextDocument>>
Metadata:  7
```

Now, we have the same number of characters, the punctuation, numbers and quotes have been removed, and everything is in lower case. Following the rubric, we proceed to compute the Document Term Matrix of the clean corpus.

```
> POM_DTM <- DocumentTermMatrix(POMLow)
> POM_DTM
<<DocumentTermMatrix (documents: 11, terms: 3866)>>
Non-/sparse entries: 8205/34321
Sparsity           : 81%
Maximal term length: 17
Weighting          : term frequency (tf)
> str(POM_DTM
+ )
List of 6
 $ i       : int [1:8205] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:8205] 2 3 18 25 30 32 35 57 59 66 ...
 $ v       : num [1:8205] 1 7 1 1 4 1 1 1 1 3 ...
 $ nrow    : int 11
 $ ncol    : int 3866
 $ dimnames:List of 2
  ..$ Docs : chr [1:11] "book_chapter1.txt" "book_chapter10.txt" "book_chapter11.txt" "book_chapter2.txt" ...
  ..$ Terms: chr [1:3866] "ability" "able" "about" "above" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
> inspect(POM_DTM)
<<DocumentTermMatrix (documents: 11, terms: 3866)>>
Non-/sparse entries: 8205/34321
Sparsity           : 81%
Maximal term length: 17
Weighting          : term frequency (tf)
Sample             :
                    Terms
Docs                 and but for had that the upon was which with
  book_chapter1.txt   93  13  20  27   50 190   12  38    20   25
  book_chapter10.txt 117  20  34  41   50 193   19  54    16   26
  book_chapter11.txt  89  14  10  29   38 122   13  19    13   23
  book_chapter2.txt   59  13  17  20   20 137   13  33    13   10
  book_chapter3.txt   94  17  18  19   26 166   20  37    34   20
  book_chapter4.txt   70  21  10  22   19 148   11  26    17   20
  book_chapter5.txt   56  14  13  14   11  96    8  22     9    8
  book_chapter6.txt   57  10  15  29   11 117   13  18    12   24
  book_chapter7.txt   67   9  18  16   12 166    7  18    24   12
  book_chapter8.txt   84   7  15  24   11 184   22  32    17   11
```

We notice our sparsity has reduced by 2% to 81%. However, this is till very high and we will employ techniques to further reduce it. Sparsity indicates the number of 0s in our Document Term Matrix. The high sparsity in our corpus can be visualized by viewing our DTM using the as.matrix(POM_DTM) method. We can observe a lot of 0s here, confirming the 81% sparsity.

```
                  Terms
Docs               would wounds wraith wriggling wring wrinkling write written wrong wrought yards year yearly
  book_chapter1.txt    13      0      0         0     0         0     1       0     2       0     1    1      0
  book_chapter10.txt   21      0      0         1     0         1     1       1     2       0     0    0      0
  book_chapter11.txt   12      0      0         0     1         0     0       1     0       2     0    0      0
  book_chapter2.txt     2      0      1         0     0         0     0       0     0       1     0    0      0
  book_chapter3.txt     7      0      0         0     0         0     0       0     0       0     3    0      0
  book_chapter4.txt     6      0      0         0     0         0     0       0     0       1     0    0      0
  book_chapter5.txt     4      0      0         0     0         0     0       0     1       1     0    0      0
  book_chapter6.txt     1      1      0         0     0         0     0       0     0       0     0    0      0
  book_chapter7.txt     4      0      0         0     0         0     0       0     0       0     0    4      2
  book_chapter8.txt     3      0      0         0     0         0     0       0     0       1     0    0      0
  book_chapter9.txt     4      0      0         0     0         0     0       0     1       1     0    0      0
                  Terms
Docs               yearning years yellow yellowish yellowishgreen yells yes yesterday yet you young younger your
  book_chapter1.txt        0     4      0         0              0     1   0         0   3   2     0       0    0
  book_chapter10.txt       0     1      0         0              0     0   1         0   1  47     3       0   24
  book_chapter11.txt       0     1      1         0              0     0   0         0   3  34     2       0   13
  book_chapter2.txt        0     1      0         0              0     0   0         0   4   0     0       0    0
  book_chapter3.txt        0     0      1         1              1     0   0         0   1   3     3       0    1
  book_chapter4.txt        0     3      0         0              0     0   0         0   1   0     1       0    0
  book_chapter5.txt        0     0      0         0              0     0   0         0   1   0     0       0    0
  book_chapter6.txt        0     0      0         0              0     0   0         0   0   0     1       0    0
  book_chapter7.txt        0     6      0         0              0     0   0         1   0   0     7       1    0
  book_chapter8.txt        1     1      0         0              0     0   0         0   3   0     0       0    0
  book_chapter9.txt        0     1      0         0              0     0   0         0   3   5     4       0    0
                  Terms
Docs               yourself yourselves youth youthful zealously
  book_chapter1.txt        0          0     0        0         0
  book_chapter10.txt       1          1     0        0         0
  book_chapter11.txt       0          0     0        0         0
  book_chapter2.txt        0          0     0        0         0
  book_chapter3.txt        0          0     0        0         0
  book_chapter4.txt        0          0     0        0         0
  book_chapter5.txt        0          0     0        0         0
  book_chapter6.txt        0          0     0        0         1
  book_chapter7.txt        0          0     1        0         0
  book_chapter8.txt        0          0     0        0         0
  book_chapter9.txt        0          0     0        1         0
```

Next we remove the stop words from our DTM.

```
> myStopwords <- c(tm::stopwords("english"))
> myStopwords
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"
  [9] "you"        "your"       "yours"      "yourself"   "yourselves" "he"         "him"        "his"
 [17] "himself"    "she"        "her"        "hers"       "herself"    "it"         "its"        "itself"
 [25] "they"       "them"       "their"      "theirs"     "themselves" "what"       "which"      "who"
 [33] "whom"       "this"       "that"       "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"      "have"       "has"        "had"
 [49] "having"     "do"         "does"       "did"        "doing"      "would"      "should"     "could"
 [57] "ought"      "i'm"        "you're"     "he's"       "she's"      "it's"       "we're"      "they're"
 [65] "i've"       "you've"     "we've"      "they've"    "i'd"        "you'd"      "he'd"       "she'd"
 [73] "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
 [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"     "haven't"    "hadn't"     "doesn't"
 [89] "don't"      "didn't"     "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"
 [97] "couldn't"   "mustn't"    "let's"      "that's"     "who's"      "what's"     "here's"     "there's"
[105] "when's"     "where's"    "why's"      "how's"      "a"          "an"         "the"        "and"
[113] "but"        "if"         "or"         "because"    "as"         "until"      "while"      "of"
[121] "at"         "by"         "for"        "with"       "about"      "against"    "between"    "into"
[129] "through"    "during"     "before"     "after"      "above"      "below"      "to"         "from"
[137] "up"         "down"       "in"         "out"        "on"         "off"        "over"       "under"
[145] "again"      "further"    "then"       "once"       "here"       "there"      "when"       "where"
[153] "why"        "how"        "all"        "any"        "both"       "each"       "few"        "more"
[161] "most"       "other"      "some"       "such"       "no"         "nor"        "not"        "only"
[169] "own"        "same"       "so"         "than"       "too"        "very"
```

```
> POMStop <- tm::tm_map(POMLow, tm::removeWords, myStopwords)
> tm::inspect(POMStop[[1]])
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 9812


  arizona hills


    old man  old    know  possibly    hundred
possibly     tell    never aged
men    remember  childhood   far   can recollect
always   man  man   thirty   appear today
forty years   ago  yet  feel    go  living
forever   day  shall die  real death
 resurrection     know    fear death
died twice   still alive  yet     horror
   never died       terror  death
believe     convinced   mortality

    conviction   determined  write
story   interesting periods   life    death
explain  phenomena  can  set     words
ordinary soldier  fortune  chronicle   strange events
befell    ten years   dead body lay undiscovered
arizona cave

  never told  story  shall mortal man see  manuscript
    passed   eternity   know   average
human mind will  believe    grasp
purpose  pilloried   public  pulpit   press
held    colossal liar     telling  simple truths
  day science will substantiate  possibly  suggestions
  gained upon mars   knowledge   can set
 chronicle will aid   earlier understanding   mysteries
  sister planet mysteries     longer mysteries


 name  john carter   better known  captain jack carter
virginia    close   civil war  found  possessed
several hundred thousand dollars confederate   captains
```

Here we have removed several stop words using the tm package and can inspect the first document to get a sense of what our corpus looks like. As we can see, the stop words have been removed and our corpus can be analyzed better.

Now that the stop words have been removed, we create the Term Document Matrix again. We also apply findFreqTerms with a lowFreq of 5.

```
> POMStopTDM <- tm::TermDocumentMatrix(POMStop)
> POMStopTDM
<<TermDocumentMatrix (terms: 3768, documents: 11)>>
Non-/sparse entries: 7405/34043
Sparsity           : 82%
Maximal term length: 17
Weighting          : term frequency (tf)
> freqTerms <- tm::findFreqTerms(POMStopTDM, lowfreq=5)
> freqTerms
  [1] "ability"      "across"       "act"          "advanced"     "affection"    "afterward"    "age"
  [8] "ages"         "aid"          "air"          "almost"       "alone"        "also"         "always"
 [15] "among"        "ancient"      "animal"       "animals"      "another"      "answered"     "answering"
 [22] "ape"          "appearance"   "appeared"     "approach"     "approached"   "approaching"  "arizona"
 [29] "arm"          "arms"         "around"       "asked"        "attempt"      "attention"    "attitude"
 [36] "audience"     "away"         "back"         "barsoom"      "battle"       "bearing"      "beast"
 [43] "beautiful"    "became"       "become"       "behind"       "believe"      "beside"       "better"
 [50] "beyond"       "blow"         "bodies"       "body"         "bottom"       "breast"       "bring"
 [57] "broke"        "brought"      "brute"        "building"     "buildings"    "call"         "came"
 [64] "can"          "captive"      "carried"      "carry"        "carter"       "catch"        "caught"
 [71] "cause"        "caused"       "cautiously"   "cavalcade"    "cave"         "ceased"       "center"
 [78] "chamber"      "chariots"     "chieftain"    "chieftains"   "children"     "city"         "clear"
 [85] "cliff"        "close"        "cold"         "color"        "come"         "common"       "community"
 [92] "conditions"   "considerable" "continued"    "conversation" "convinced"    "council"      "course"
 [99] "craft"        "creature"     "creatures"    "creeping"     "crude"        "cruel"        "cudgel"
```

```
> length(freqTerms) > freqTerms[5]   > nchar(freqTerms[5])
[1] 551             [1] "affection"  [1] 9
```

We check the number of our freqTerms which is 551 and can check specific frequent terms. For example, as shown the fifth term above has a length of 9. We can compute the term frequencies for each document as follows. Later on we will unlist tfList and combine all the term frequencies.

```
> tfList <- list()
> for (i in seq_along(POMStop)) {
+     tfList[[i]] <- termFreq(POMStop[[i]])
+ }
> print(tfList[[1]])
          able        account     accurately         across          acted           acts         advent     adventures
             1              1              1              4              1              1              1              1
     afternoon           aged            ago         agreed          ahead            aid            aim          alive
             1              1              1              1              2              1              1              2
        almost          alone        already    alternative         always          among        animals        another
             1              1              1              1              3              1              2              1
      antelope       anything        apaches     apartments         appear   apprehension        arizona            arm
             1              1              3              1              1              1              5              1
         armed         arming           army         arose         around         arrows         assure       attacked
             1              1              3              1              3              3              1              1
       attempt      attention     attraction     attributed        average          await           back       backward
             1              1              1              1              1              1              3              1
         balls         bathed         beast      beautiful         became         become         befell        believe
             1              1              1              1              1              2              2              3
          belt          belts           best       bestowed         better        bidding           body         borne
             1              1              1              1              3              1              5              1
        bottom           bows         braves          brief         bright          brisk      bristling          broad
             1              1              1              1              1              1              1              1
         broke        brought         burros           came           camp            can        canteen         canter
             1              1              1              3              3              3              1              1
```

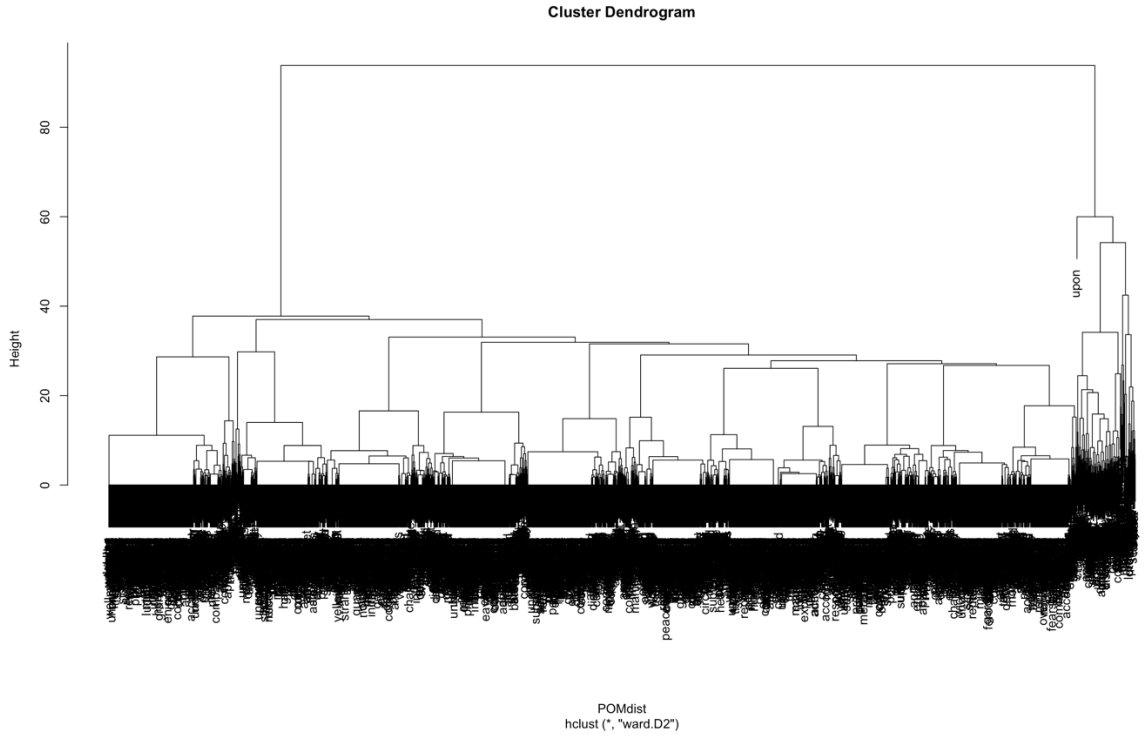
We can inspect POMStopTDM again

```
> tm::inspect(POMStopTDM)
<<TermDocumentMatrix (terms: 3768, documents: 11)>>
Non-/sparse entries: 7405/34043
Sparsity           : 82%
Maximal term length: 17
Weighting          : term frequency (tf)
Sample             :
          Docs
Terms      book_chapter1.txt book_chapter10.txt book_chapter11.txt book_chapter2.txt book_chapter3.txt
  feet                     5                 5                  0                 1                15
  first                    3                 6                  3                 5                 5
  little                   4                 4                  4                 2                11
  mars                     1                 4                  2                 1                11
  martian                  0                13                  3                 0                 6
  martians                 0                 1                  4                 0                 5
  one                      3                19                  1                 2                 7
  sola                     0                 6                 11                 0                 0
  toward                   3                 9                  1                 3                10
  upon                    12                19                 13                13                20
          Docs
Terms      book_chapter4.txt book_chapter5.txt book_chapter6.txt book_chapter7.txt book_chapter8.txt
  feet                     9                 2                  4                 2                 1
  first                    6                 1                  3                 3                 5
  little                   3                 2                  0                10                 4
  mars                     6                 9                  2                 4                 1
  martian                  9                 5                  2                10                 7
  martians                 8                 1                  7                11                 3
  one                      5                 6                  5                 9                 8
  sola                     2                 5                  4                 7                 3
  toward                   6                 3                  3                 2                 5
  upon                    11                 8                 13                 7                22
```

5. Dendrogram and Word Cloud

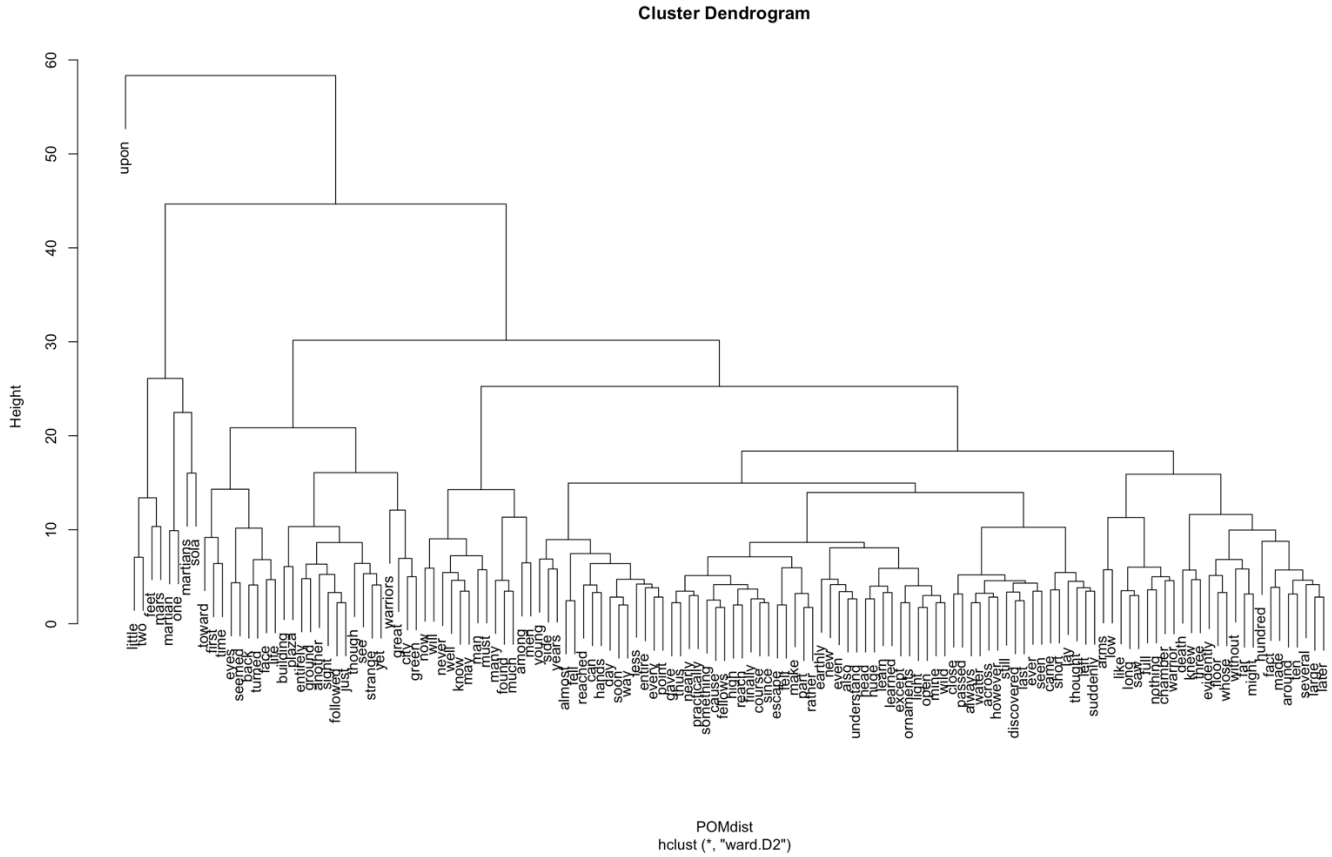
Next, we can draw a dendrogram for POMStopTDM.

```
> POMdf<- as.data.frame(as.matrix(POMStopTDM))
> POMdist<- dist(POMdf)
> POMDG<- hclust(POMdist, method="ward.D2")
> str(POMDG)
List of 7
 $ merge      : int [1:3767, 1:2] -2 -1085 -2647 -3 -40 -81 -82 -84 -154 -162 ...
 $ height     : num [1:3767] 0 0 0 0 0 0 0 0 0 0 ...
 $ order      : int [1:3768] 3761 3747 3745 3742 3730 3712 3677 3660 3628 3616 ...
 $ labels     : chr [1:3768] "ability" "able" "abreast" "abruptly" ...
 $ method     : chr "ward.D2"
 $ call       : language hclust(d = POMdist, method = "ward.D2")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> plot(POMDG)
```



We can observe that we have many labels, and it is hard to analyze our dendrogram. Therefore, it would be helpful to remove some of the sparse terms to get a good dendrogram. We utilized the removeSparseTerms() function from the tm package to remove the sparse terms from the term document matrix. We experimented with different values for maximal allowed sparsity and found that a value of 0.4 yields us a comprehensive dendrogram that can be analyzed. We get a sparsity of 24%. (We did not prefer reducing the sparsity to 0% as that would lose a great percentage of the terms and we want as many relevant terms as possible for our analysis).

```
> POMStopTDM_rmSparse <- removeSparseTerms(POMStopTDM, 0.4)
> tm::inspect(POMStopTDM_rmSparse)
<<TermDocumentMatrix (terms: 137, documents: 11)>>
Non-/sparse entries: 1146/361
Sparsity           : 24%
Maximal term length: 11
Weighting          : term frequency (tf)
Sample             :
          Docs
Terms      book_chapter1.txt book_chapter10.txt book_chapter11.txt book_chapter3.txt
  feet                     5                  5                  0                15
  first                    3                  6                  3                 5
  little                   4                  4                  4                11
  mars                     1                  4                  2                11
  martian                  0                 13                  3                 6
  martians                 0                  1                  4                 5
  one                      3                 19                  1                 7
  sola                     0                  6                 11                 0
  toward                   3                  9                  1                10
  upon                    12                 19                 13                20
          Docs
Terms      book_chapter4.txt book_chapter5.txt book_chapter6.txt book_chapter7.txt
  feet                     9                  2                  4                 2
  first                    6                  1                  3                 3
  little                   3                  2                  0                10
  mars                     6                  9                  2                 4
  martian                  9                  5                  2                10
  martians                 8                  1                  7                11
  one                      5                  6                  5                 9
  sola                     2                  5                  4                 7
  toward                   6                  3                  3                 2
  upon                    11                  8                 13                 7
```



We can now proceed to generate a word cloud. We use the wordcloud package for this, install it, and make it a library.

```
> install.packages("wordcloud")
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/wordcloud_2.6.tgz'
Content type 'application/x-gzip' length 276045 bytes (269 KB)
==================================================
downloaded 269 KB


The downloaded binary packages are in
        /var/folders/d5/w02vn_rj1hv2m765h0gj7qdc0000gn/T//RtmpHuXVVD/downloaded_packages
> library(wordcloud)
Loading required package: RColorBrewer
```

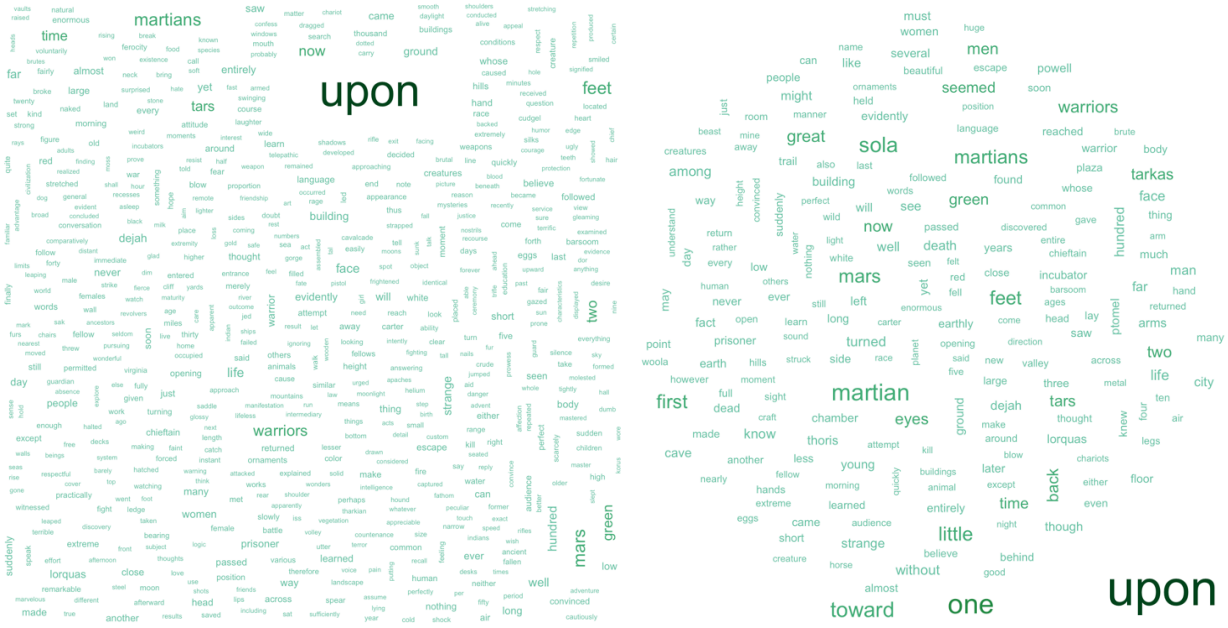
```
> tf_combined <- unlist(tfList)
> tf_summed <- tapply(tf_combined, names(tf_combined), sum)
> words<- names(tf_summed)
> words
   [1] "ability"         "able"            "abreast"         "abruptly"        "absence"         "absolutely"
   [7] "abundantly"      "accompanied"     "accompany"       "accomplished"    "accomplishing"   "accord"
  [13] "accordance"      "accorded"        "according"       "account"         "accounted"       "accounting"
  [19] "accounts"        "accouterments"   "accoutrements"   "accuracy"        "accurately"      "accustomed"
  [25] "acid"            "acquainted"      "acquired"        "across"          "act"             "acted"
  [31] "action"          "actions"         "acts"            "actual"          "adapted"         "add"
  [37] "added"           "adding"          "addition"        "additional"      "address"         "addressed"
  [43] "addressing"      "adjoining"       "administration"  "ado"             "adoration"       "adult"
  [49] "adults"          "advance"         "advanced"        "advancement"     "advancing"       "advantage"
  [55] "advent"          "adventure"       "adventures"      "adversary"       "advising"        "affect"
  [61] "affection"       "affections"      "african"         "afternoon"       "afterward"       "afterwards"
  [67] "age"             "aged"            "ageold"          "ages"            "agile"           "agility"
  [73] "ago"             "agonies"         "agreed"          "ahead"           "aid"             "aim"
  [79] "aimed"           "air"             "airtight"        "aisle"           "albinos"         "alien"
  [85] "alighted"        "alighting"       "alike"           "alive"           "alliance"        "allowed"
  [91] "allowing"        "alloy"           "almost"          "alone"           "along"           "aloud"
  [97] "already"         "also"            "alternative"     "although"        "altogether"      "aluminum"
 [103] "always"          "america"         "amity"           "ammunition"      "among"           "amongst"
 [109] "amounted"        "amusement"       "anaesthesis"     "ancestor"        "ancestors"       "ancient"
 [115] "andi"            "angrily"         "animal"          "animals"         "annihilate"      "annihilation"
 [121] "anomalies"       "another"         "answered"        "answering"       "antagonist"      "antecedents"
 [127] "antelope"        "antennae"        "antennaelike"    "antics"          "antiquity"       "anxious"
 [133] "anxiously"       "anything"        "anyway"          "anywhere"        "apache"          "apaches"
 [139] "apartments"      "ape"             "apelike"         "apes"            "appalled"        "appalling"
 [145] "apparatus"       "apparel"         "apparent"        "apparently"      "appeal"          "appear"
 [151] "appearance"      "appeared"        "applause"        "appraised"       "appreciable"     "apprehension"
 [157] "apprised"        "approach"        "approached"      "approaching"     "approbation"     "appropriated"
 [163] "apt"             "arc"             "architects"      "architecture"    "archives"        "area"
 [169] "areas"           "argument"        "arid"            "arises"          "arising"         "arizona"
 [175] "arm"             "armada"          "armed"           "arming"          "armlet"          "arms"
 [181] "army"            "arose"           "around"          "arouse"          "aroused"         "arrange"
 [187] "array"           "arrived"         "arrows"          "arroyo"          "art"             "article"
 [193] "artificial"      "artisans"        "arts"            "ascendency"      "ascending"       "ask"
 [199] "asked"           "asking"          "asleep"          "aspect"          "assemblage"      "assembled"
 [205] "assistance"      "assisting"       "assume"          "assure"          "astir"           "astonishment"
 [211] "astute"          "atavism"         "ate"             "atmosphere"      "atmospheric"     "atrophied"
 [217] "attack"          "attacked"        "attacking"       "attempt"         "attempted"       "attempting"
 [223] "attempts"        "attend"          "attends"         "attention"       "attentions"      "attitude"
```

Here we extract the unique words using the term frequencies for all the 11 documents. We unlist the tfList variable to combine all the term frequencies, then we sum up the term frequencies for each unique term. And then we obtain the set of words into the **words** variable.

Next, we generate the wordcloud using the wordcloud() function. We pass the summed term frequencies and the words to generate it. Some warnings occurred and upon examining them, we found out that some words were too long to fit in the wordcloud and were thus omitted. To tackle this, we generate another wordcloud, keeping min.freq = 10 to only display words that appear at least 10 times in the corpus.

```
> pal <- brewer.pal(9, "BuGn")
> str(pal)
 chr [1:9] "#F7FCFD" "#E5F5F9" "#CCECE6" "#99D8C9" "#66C2A4" "#41AE76" "#238B45" "#006D2C" "#00441B"
> POMWC <- wordcloud(words, tf_summed, colors=pal[-(1:4)])
There were 50 or more warnings (use warnings() to see the first 50)
```

```
POMWC <- wordcloud(words, tf_summed, min.freq = 10, colors=pal[-(1:4)])
```



The second word cloud helps us understand the word cloud a bit better.

5. Applying functions from packages to the Corpus

Following the rubric, we go through several packages like “quanteda”, “syuzhet”, etc. to perform further analysis.

```
> install.packages("quanteda")
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/quanteda_3.3.0.tgz'
Content type 'application/x-gzip' length 4294524 bytes (4.1 MB)
==================================================
downloaded 4.1 MB


The downloaded binary packages are in
        /var/folders/d5/w02vn_rj1hv2m765h0gj7qdc0000gn/T//RtmpHuXVVD/downloaded_packages
> library(quanteda)
Package version: 3.3.0
Unicode version: 14.0
ICU version: 71.1
Parallel computing: 8 of 8 threads used.
See https://quanteda.io for tutorials and examples.

Attaching package: 'quanteda'

The following object is masked from 'package:tm':

    stopwords

The following objects are masked from 'package:NLP':

    meta, meta<-
```

We use the clean copy of POMcl to perform further analysis. Printing out the first few lines of the first document:

```
> POMtext<- POMcl[[1]]
> POMtext$content[1:10]
 [1] ""
 [2] "ON THE ARIZONA HILLS"
 [3] ""
 [4] ""
 [5] "I am a very old man how old I do not know  Possibly I am a hundred"
 [6] "possibly more but I cannot tell because I have never aged as other"
 [7] "men nor do I remember any childhood  So far as I can recollect I have"
 [8] "always been a man a man of about thirty  I appear today as I did"
 [9] "forty years and more ago and yet I feel that I cannot go on living"
[10] "forever that some day I shall die the real death from which there is"
```

Let us apply a few methods from the quanteda package on our corpus. Kwic() helps locate keywords in context from the text.

```
> kwic(POMTokens[[1]], pattern="mars")
Keyword-in-context with 1 match.
 [text29, 5] which I gained upon | Mars | and the knowledge which I

> kwic(POMTokens[[2]], pattern="mars")
Keyword-in-context with 4 matches.
  [text79, 8]        a few peaks on all | Mars | exceed four thousand feet in
 [text98, 11] know that someone else on | Mars | beside
 [text163, 1]                           | Mars | to support a single human
 [text187, 4]           mighty era for | Mars |

> kwic(POMTokens[[3]], pattern="mars")
Keyword-in-context with 2 matches.
 [text14, 13]   do not kill women upon | Mars | nor
 [text124, 6] Barsoom which we know as | Mars | How I came here I

> kwic(POMTokens[[4]], pattern="mars")
Keyword-in-context with 1 match.
 [text153, 8] spell of overpowering fascinationit was | Mars | the god of war
```

Next we apply tokenization. We do this by using the lapply() function to apply tokenization to each of the 11 documents of the corpus.

```
> POMTokens <- lapply(POMcl, function(x) quanteda::tokens(x$content))
```

If we examine the structure of POMTokens using str(), we see a huge output for all the eleven documents.

```
  ..$ text80 : chr(0)
  ..$ text81 : chr [1:12] "Sarkoja" "one" "of" "the" ...
  ..$ text82 : chr [1:13] "present" "at" "the" "audience" ...
  ..$ text83 : chr [1:5] "toward" "her" "the" "question" ...
  ..$ text84 : chr(0)
  ..$ text85 : chr [1:14] "When" "asked" "one" "of" ...
  ..$ text86 : chr [1:12] "red" "one" "or" "does" ...
  ..$ text87 : chr(0)
  ..$ text88 : chr [1:14] "They" "have" "decided" "to" ...
  ..$ text89 : chr [1:11] "last" "agonies" "at" "the" ...
  ..$ text90 : chr(0)
  ..$ text91 : chr [1:13] "What" "will" "be" "the" ...
  ..$ text92 : chr [1:14] "very" "small" "and" "very" ...
  ..$ text93 : chr "ransom"
  ..$ text94 : chr(0)
  ..$ text95 : chr [1:11] "Sarkoja" "and" "the" "other" ...
  ..$ text96 : chr [1:6] "weakness" "on" "the" "part" ...
  ..$ text97 : chr(0)
  ..$ text98 : chr [1:14] "It" "is" "sad" "Sola" ...
  ..$ text99 : chr [1:13] "Sarkoja" "when" "all" "the" ...
  .. [list output truncated]
  ..- attr(*, "types")= chr [1:583] "I" "LEARN" "THE" "LANGUAGE" ...
  ..- attr(*, "padding")= logi FALSE
  ..- attr(*, "class")= chr "tokens"
  ..- attr(*, "docvars")='data.frame':  148 obs. of  3 variables:
  .. ..$ docname_: chr [1:148] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 148 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:148] 1 1 1 1 1 1 1 1 1 1 ...
  ..- attr(*, "meta")=List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 3 3 0
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 2 2
  .. .. ..$ system         : Named chr [1:3] "Darwin" "arm64" "amishraj"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "/Users/amishraj/Documents/Study/Big Data Analytics/Class Project 3/R Projects/APrincessOfMar
s"
  .. .. ..$ created        : Date[1:1], format: "2023-05-05"
  .. ..$ object:List of 6
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ summary     :List of 2
  .. .. .. ..$ hash: chr(0)
  .. .. .. ..$ data: NULL
  .. ..$ user  : list()
```

We now use the dfm() function to create sparse document-feature matrices for all the chapters. Here, we also remove the stopwords using dfm_remove()

```
> POMDFMCh1<- quanteda::dfm(POMTokens[[1]]) %>% dfm_remove(myStopwords)
> POMDFMCh2<- quanteda::dfm(POMTokens[[2]]) %>% dfm_remove(myStopwords)
> POMDFMCh3<- quanteda::dfm(POMTokens[[3]]) %>% dfm_remove(myStopwords)
> POMDFMCh4<- quanteda::dfm(POMTokens[[4]]) %>% dfm_remove(myStopwords)
> POMDFMCh5<- quanteda::dfm(POMTokens[[5]]) %>% dfm_remove(myStopwords)
> POMDFMCh6<- quanteda::dfm(POMTokens[[6]]) %>% dfm_remove(myStopwords)
> POMDFMCh7<- quanteda::dfm(POMTokens[[7]]) %>% dfm_remove(myStopwords)
> POMDFMCh8<- quanteda::dfm(POMTokens[[8]]) %>% dfm_remove(myStopwords)
> POMDFMCh9<- quanteda::dfm(POMTokens[[9]]) %>% dfm_remove(myStopwords)
> POMDFMCh10<- quanteda::dfm(POMTokens[[10]]) %>% dfm_remove(myStopwords)
> POMDFMCh11<- quanteda::dfm(POMTokens[[11]]) %>% dfm_remove(myStopwords)
>
> str(POMDFMCh2)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    358 obs. of  3 variables:
  .. ..$ docname_: chr [1:358] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 358 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:358] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 3 3 0
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 2 2
  .. .. ..$ system         : Named chr [1:3] "Darwin" "arm64" "amishraj"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "/Users/amishraj/Documents/Study/Big Data Analytics/Class Project 3/R Projects/APrincessOfMar
s"
  .. .. ..$ created        : Date[1:1], format: "2023-05-05"
  .. ..$ object:List of 9
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
```

Next we get the frequency of terms in the DFM for the chapters.

```
> POMDocFreq_Ch1<- quanteda::docfreq(POMDFMCh1)
> POMDocFreq_Ch2<- quanteda::docfreq(POMDFMCh2)
> POMDocFreq_Ch3<- quanteda::docfreq(POMDFMCh3)
> POMDocFreq_Ch4<- quanteda::docfreq(POMDFMCh4)
> POMDocFreq_Ch5<- quanteda::docfreq(POMDFMCh5)
> POMDocFreq_Ch6<- quanteda::docfreq(POMDFMCh6)
> POMDocFreq_Ch7<- quanteda::docfreq(POMDFMCh7)
> POMDocFreq_Ch8<- quanteda::docfreq(POMDFMCh8)
> POMDocFreq_Ch9<- quanteda::docfreq(POMDFMCh9)
> POMDocFreq_Ch10<- quanteda::docfreq(POMDFMCh10)
> POMDocFreq_Ch11<- quanteda::docfreq(POMDFMCh11)
>
> str(POMDocFreq_Ch5)
 Named int [1:773] 1 11 1 6 19 2 1 2 1 2 ...
 - attr(*, "names")= chr [1:773] "advent" "mars" "opened" "eyes" ...
> POMDocFreq_Ch5
        advent           mars         opened           eyes           upon        strange          weird
             1             11              1              6             19              2              1
     landscape           knew       question         either         sanity     wakefulness         asleep
             2              1              2              4              2              1              1
          need       pinching          inner   consciousness           told        plainly      conscious
             1              1              1              1              1              1              1
          mind          tells          earth           fact        neither          found          lying
             2              1             11              3              2              2              2
         prone            bed      yellowish       mosslike     vegetation      stretched         around
             1              1              1              1              2              1              2
    directions   interminable          miles         seemed           deep       circular          basin
             2              1              3              5              1              1              1
         along          outer         verge     distinguish  irregularities            low          hills
             2              1              1              1              1              8              2
        midday            sun        shining           full           heat         rather        intense
             1              1              1              1              1              2              1
         naked           body            yet        greater           true        similar     conditions
             3              1              1              1              1              3              1
       arizona         desert         slight    outcroppings   quartzbearing           rock      glistened
             1              1              1              1              1              1              1
      sunlight         little           left        perhaps        hundred          yards       appeared
             2             11              2              2              7              3              2
        walled      enclosure           four           feet         height          water           moss
             1              6              5             12              3              1              3
      evidence       somewhat        thirsty     determined      exploring      springing       received
             2              1              1              3              1              1              1
         first        martian       surprise         effort        brought       standing        upright
             5              6              2              2              1              1              1
       carried            air          three       alighted         softly         ground        however
             4              3              3              2              1              3              2
```

We now assign weights to these words.

```
> POMWeights_Ch1 <- quanteda::dfm_weight(POMDFMCh1)
> POMWeights_Ch2 <- quanteda::dfm_weight(POMDFMCh2)
> POMWeights_Ch3 <- quanteda::dfm_weight(POMDFMCh3)
> POMWeights_Ch4 <- quanteda::dfm_weight(POMDFMCh4)
> POMWeights_Ch5 <- quanteda::dfm_weight(POMDFMCh5)
> POMWeights_Ch6 <- quanteda::dfm_weight(POMDFMCh6)
> POMWeights_Ch7 <- quanteda::dfm_weight(POMDFMCh7)
> POMWeights_Ch8 <- quanteda::dfm_weight(POMDFMCh8)
> POMWeights_Ch9 <- quanteda::dfm_weight(POMDFMCh9)
> POMWeights_Ch10 <- quanteda::dfm_weight(POMDFMCh10)
> POMWeights_Ch11 <- quanteda::dfm_weight(POMDFMCh11)
> str(POMWeights_Ch1)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    267 obs. of  3 variables:
  .. ..$ docname_: chr [1:267] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 267 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:267] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 3 3 0
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 2 2
  .. .. ..$ system         : Named chr [1:3] "Darwin" "arm64" "amishraj"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "/Users/amishraj/Documents/Study/Big Data Analytics/Class Project 3/R Projects/APrincessOfMar
s"
  .. .. ..$ created        : Date[1:1], format: "2023-05-05"
  .. ..$ object:List of 9
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ ngram       : int 1
```

```
> POMWeights_Ch1
Document-feature matrix of: 267 documents, 788 features (99.43% sparse) and 0 docvars.
       features
docs    arizona hills old man know possibly hundred tell never aged
  text1       0     0   0   0    0        0       0    0     0    0
  text2       1     1   0   0    0        0       0    0     0    0
  text3       0     0   0   0    0        0       0    0     0    0
  text4       0     0   0   0    0        0       0    0     0    0
  text5       0     0   2   1    1        1       1    0     0    0
  text6       0     0   0   0    0        1       0    1     1    1
[ reached max_ndoc ... 261 more documents, reached max_nfeat ... 778 more features ]
```

Now we compute the tf-idf score. This helps us weigh a dfm by term frequency-inverse document frequency (tf-idf), with full control over options.

```
> POMTFIDF_Ch1<- quanteda::dfm_tfidf(POMDFMCh1, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch2<- quanteda::dfm_tfidf(POMDFMCh2, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch3<- quanteda::dfm_tfidf(POMDFMCh3, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch4<- quanteda::dfm_tfidf(POMDFMCh4, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch5<- quanteda::dfm_tfidf(POMDFMCh5, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch6<- quanteda::dfm_tfidf(POMDFMCh6, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch7<- quanteda::dfm_tfidf(POMDFMCh7, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch8<- quanteda::dfm_tfidf(POMDFMCh8, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch9<- quanteda::dfm_tfidf(POMDFMCh9, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch10<- quanteda::dfm_tfidf(POMDFMCh10, scheme_tf = "count", scheme_df = "inverse")
> POMTFIDF_Ch11<- quanteda::dfm_tfidf(POMDFMCh11, scheme_tf = "count", scheme_df = "inverse")
```

```
> str(POMTFIDF_Ch1)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    267 obs. of  3 variables:
  .. ..$ docname_: chr [1:267] "text1" "text2" "text3" "text4" ...
  .. ..$ docid_  : Factor w/ 267 levels "text1","text2",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:267] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 3 3 0
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 4 2 2
  .. .. ..$ system         : Named chr [1:3] "Darwin" "arm64" "amishraj"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "/Users/amishraj/Documents/Study/Big Data Analytics/Class Project 3/R Projects/APrincessOfMar
s"
  .. .. ..$ created        : Date[1:1], format: "2023-05-05"
  .. ..$ object:List of 9
  .. .. ..$ unit        : chr "documents"
  .. .. ..$ what        : chr "word"
  .. .. ..$ ngram       : int 1
  .. .. ..$ skip        : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf   :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
  .. .. .. ..$ k     : NULL
  .. .. ..$ weight_df   :List of 2
  .. .. .. ..$ scheme: chr "inverse"
  .. .. .. ..$ base  : num 10
  .. .. ..$ smooth      : num 0
  .. .. ..$ summary     :List of 2
  .. .. .. ..$ hash: chr(0)
  .. .. .. ..$ data: NULL
  .. ..$ user  : list()
  ..@ i       : int [1:1197] 1 20 63 170 228 1 4 103 108 4 ...
  ..@ p       : int [1:789] 0 5 6 9 15 19 25 30 31 36 ...
  ..@ Dim     : int [1:2] 267 788
  ..@ Dimnames:List of 2
  .. ..$ docs    : chr [1:267] "text1" "text2" "text3" "text4" ...
  .. ..$ features: chr [1:788] "arizona" "hills" "old" "man" ...
  ..@ x       : Named num [1:1197] 1.73 1.73 1.73 1.73 1.73 ...
  .. ..- attr(*, "names")= chr [1:1197] "arizona" "arizona" "arizona" "arizona" ...
  ..@ factors : list()
```

Now we install the “syuzhet” package and make it a library.

```
> install.packages("syuzhet")
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/syuzhet_1.0.6.tgz'
Content type 'application/x-gzip' length 3114882 bytes (3.0 MB)
==================================================
downloaded 3.0 MB


The downloaded binary packages are in
        /var/folders/d5/w02vn_rj1hv2m765h0gj7qdc0000gn/T//RtmpHuXVVD/downloaded_packages
> library(syuzhet)
```

We can extract text as a data frame. Let us apply a few functions from the syuzhet package on our document.

```
> POMTextCh1<- POMcl[[1]]
> POMTextCh1$content[1:10]
 [1] ""
 [2] "ON THE ARIZONA HILLS"
 [3] ""
 [4] ""
 [5] "I am a very old man how old I do not know  Possibly I am a hundred"
 [6] "possibly more but I cannot tell because I have never aged as other"
 [7] "men nor do I remember any childhood  So far as I can recollect I have"
 [8] "always been a man a man of about thirty  I appear today as I did"
 [9] "forty years and more ago and yet I feel that I cannot go on living"
[10] "forever that some day I shall die the real death from which there is"
>
> POMTextdf<- as.data.frame(POMcl[[1]]$content)
> POMTextdf
                                                    POMcl[[1]]$content
1
2                                                  ON THE ARIZONA HILLS
3
4
5        I am a very old man how old I do not know  Possibly I am a hundred
6        possibly more but I cannot tell because I have never aged as other
7     men nor do I remember any childhood  So far as I can recollect I have
8          always been a man a man of about thirty  I appear today as I did
9        forty years and more ago and yet I feel that I cannot go on living
10    forever that some day I shall die the real death from which there is
11       no resurrection  I do not know why I should fear death I who have
12    died twice and am still alive but yet I have the same horror of it as
13     you who have never died and it is because of this terror of death I
14                          believe that I am so convinced of my mortality
```

To get the sentences in APrincessOfMars.txt, we first read the file as one large string as shown with the help of get_text_as_string() function.

```
> POMAsString<- get_text_as_string("APrincessOfMars.txt")
> POMAsString
 CHAPTER I  ON THE ARIZONA HILLS   I am a very old man; how old I do not know.  Possibly I am a hundred, possibly more; but I
cannot tell because I have never aged as other men, nor do I remember any childhood.  So far as I can recollect I have always
been a man, a man of about thirty.  I appear today as I did forty years and more ago, and yet I feel that I cannot go on livi
g forever; that some day I shall die the real death from which there is no resurrection.  I do not know why I should fear deat
h, I who have died twice and am still alive; but yet I have the same horror of it as you who have never died, and it is becaus
e of this terror of death, I believe, that I am so convinced of my mortality.  And because of this conviction I have determin
d to write down the story of the interesting periods of my life and of my death.  I cannot explain the phenomena; I can only s
et down here in the words of an ordinary soldier of fortune a chronicle of the strange events that befell me during the ten ye
ars that my dead body lay undiscovered in an Arizona cave.  I have never told this story, nor shall mortal man see this manus
ript until after I have passed over for eternity.  I know that the average human mind will not believe what it cannot grasp, a
nd so I do not purpose being pilloried by the public, the pulpit, and the press, and held up as a colossal liar when I am but
telling the simple truths which some day science will substantiate.  Possibly the suggestions which I gained upon Mars, and th
e knowledge which I can set down in this chronicle, will aid in an earlier understanding of the mysteries of our sister plane
t; mysteries to you, but no longer mysteries to me.  My name is John Carter; I am better known as Captain Jack Carter of Virg
```

Now we get the sentences using the get_sentences() function from the syuzhet package. It parses the string into individual sentences.

```
> POMS<- get_sentences(POMAsString)
> POMS[1:10]
 [1] "CHAPTER I  ON THE ARIZONA HILLS   I am a very old man; how old I do not know."
 [2] "Possibly I am a hundred, possibly more; but I cannot tell because I have never aged as other men, nor do I remember any
childhood."
 [3] "So far as I can recollect I have always been a man, a man of about thirty."
 [4] "I appear today as I did forty years and more ago, and yet I feel that I cannot go on living forever; that some day I sha
ll die the real death from which there is no resurrection."
 [5] "I do not know why I should fear death, I who have died twice and am still alive; but yet I have the same horror of it as
you who have never died, and it is because of this terror of death, I believe, that I am so convinced of my mortality."
 [6] "And because of this conviction I have determined to write down the story of the interesting periods of my life and of my
death."
 [7] "I cannot explain the phenomena; I can only set down here in the words of an ordinary soldier of fortune a chronicle of t
he strange events that befell me during the ten years that my dead body lay undiscovered in an Arizona cave."
 [8] "I have never told this story, nor shall mortal man see this manuscript until after I have passed over for eternity."
 [9] "I know that the average human mind will not believe what it cannot grasp, and so I do not purpose being pilloried by the
public, the pulpit, and the press, and held up as a colossal liar when I am but telling the simple truths which some day scien
ce will substantiate."
[10] "Possibly the suggestions which I gained upon Mars, and the knowledge which I can set down in this chronicle, will aid in
an earlier understanding of the mysteries of our sister planet; mysteries to you, but no longer mysteries to me."
```

We now proceed to extract the sentiments from these sentences. To do this, we utilize the get_sentiment() method that takes a vector of strings and one of four methods for sentiment extraction, "syuzhet", "bing", "afinn", and "nrc". We experiment with all of these methods to observe the results. POMS is a character vector.

```
> str(POMS)
 chr [1:2321] "CHAPTER I  ON THE ARIZONA HILLS   I am a very old man; how old I do not know." ...
```

We first call get_sentiment() with the default value of "syuzhet"

```
> POMSSentiment<- get_sentiment(POMS, "syuzhet")
> POMSSentiment
   [1]  0.000000e+00  6.000000e-01  0.000000e+00 -1.000000e+00 -3.400000e+00  0.000000e+00 -9.000000e-01 -2.500000e-01
   [9] -3.000000e-01  3.200000e+00  8.000000e-01 -1.000000e-01 -1.050000e+00 -6.500000e-01  9.500000e-01  6.500000e-01
  [17]  2.000000e+00  2.600000e+00  5.000000e-01  8.000000e-01  1.400000e+00  0.000000e+00  8.000000e-01  5.000000e-01
  [25] -4.950000e+00  1.650000e+00  6.500000e-01  1.000000e+00  0.000000e+00 -7.500000e-01 -2.500000e-01  2.850000e+00
  [33]  1.500000e+00 -8.500000e-01  7.000000e-01 -1.500000e+00 -5.000000e-01 -7.500000e-01  8.000000e-01 -1.500000e+00
  [41]  1.050000e+00  1.200000e+00  1.800000e+00  5.000000e-02 -1.000000e+00 -1.250000e+00  2.500000e-01  7.000000e-01
  [49] -2.000000e+00  2.500000e-01 -1.350000e+00 -6.500000e-01 -1.250000e+00 -1.250000e+00 -2.500000e-01  1.300000e+00
  [57]  1.000000e+00 -3.500000e-01  2.000000e-01 -6.000000e-01  3.000000e-01  1.000000e+00  1.200000e+00  8.000000e-01
  [65]  8.000000e-01  0.000000e+00 -1.000000e+00 -4.000000e-01 -4.500000e-01  3.700000e+00 -6.000000e-01  1.200000e+00
  [73] -7.500000e-01 -9.500000e-01  1.250000e+00  5.000000e-02 -8.500000e-01  6.000000e-01 -1.500000e+00 -5.000000e-01
  [81]  0.000000e+00 -4.000000e-01 -9.000000e-01  4.000000e-01 -2.000000e-01  1.550000e+00 -1.150000e+00 -1.500000e-01
  [89]  0.000000e+00 -5.000000e-01  6.500000e-01  0.000000e+00 -2.050000e+00 -1.600000e+00 -1.250000e+00 -5.000000e-01
  [97] -2.750000e+00 -4.700000e+00 -1.050000e+00 -5.000000e-01 -1.050000e+00 -4.900000e+00 -2.750000e+00  9.500000e-01
 [105] -1.250000e+00 -1.550000e+00 -6.500000e-01 -1.250000e+00 -7.500000e-01  0.000000e+00  4.000000e-01 -2.500000e-01
 [113] -1.000000e+00 -2.000000e-01 -1.500000e-01  2.500000e-01 -2.600000e+00 -1.500000e+00  2.150000e+00 -1.250000e+00
 [121]  3.850000e+00  2.250000e+00  3.250000e+00 -1.250000e+00  1.700000e+00  1.100000e+00 -1.350000e+00 -4.500000e-01
 [129] -7.500000e-01 -1.250000e+00  5.000000e-02  0.000000e+00  4.000000e-01  1.000000e-01 -5.000000e-01  1.450000e+00
 [137]  0.000000e+00  8.500000e-01  1.800000e+00  5.000000e-02 -7.500000e-01  2.000000e+00  1.850000e+00 -1.600000e+00
 [145]  2.500000e-01  1.300000e+00  2.500000e-01  1.150000e+00  0.000000e+00 -1.750000e+00  0.000000e+00  1.300000e+00
 [153]  4.000000e-01  2.500000e-01  0.000000e+00  6.000000e-01  4.000000e-01 -6.000000e-01  4.000000e-01 -2.050000e+00
 [161]  4.000000e-01 -4.500000e-01  1.750000e+00 -1.500000e+00  4.500000e-01 -7.000000e-01  7.500000e-01 -1.200000e+00
 [169] -1.400000e+00 -1.550000e+00  4.000000e-01  2.200000e+00  0.000000e+00  0.000000e+00  1.000000e-01  4.000000e-01
 [177]  9.000000e-01  0.000000e+00 -6.000000e-01  8.500000e-01 -6.500000e-01 -4.000000e-01  1.000000e+00  1.850000e+00
 [185] -4.000000e-01  1.900000e+00  0.000000e+00  1.000000e+00  6.000000e-01  1.950000e+00  2.850000e+00  1.250000e+00
 [193]  2.000000e-01  4.500000e-01 -1.000000e-01  1.500000e+00 -1.500000e+00  1.450000e+00 -6.000000e-01 -6.000000e-01
 [201] -6.000000e-01  7.000000e-01 -1.400000e+00 -1.000000e-01 -4.000000e-01 -2.500000e-01  1.550000e+00  1.750000e+00
 [209]  5.000000e-01  1.550000e+00  7.500000e-01  8.000000e-01 -2.500000e-01 -5.000000e-01  0.000000e+00  9.000000e-01
 [217]  0.000000e+00 -7.500000e-01  5.500000e-01  0.000000e+00 -7.500000e-01 -1.100000e+00 -3.500000e-01  0.000000e+00
 [225]  0.000000e+00 -1.100000e+00  0.000000e+00  1.200000e+00  6.000000e-01 -2.550000e+00 -2.750000e+00 -2.500000e-01
 [233]  1.950000e+00  2.500000e-01 -6.000000e-01  0.000000e+00  1.000000e-01  1.900000e+00  1.000000e+00  1.100000e+00
 [241]  2.500000e-01  1.300000e+00  1.750000e+00  2.500000e-01 -2.000000e+00  3.500000e-01  0.000000e+00  2.500000e-01
```

If we try the "bing" method, we get the following results.

```
> POMSBing<- get_sentiment(POMS, "bing")
> POMSBing
   [1]  0  0  0 -2 -4  0 -2  0 -1  1  1  0  1  0  2  1  2  2  0  2  3  0  0 -1 -8  1  0  0  0  0 -1  2  3 -1  0 -2 -1  0
  [39] -1 -2 -1  0  1 -1 -1 -1  2  1 -2  1 -2  0 -3  0 -3  0  1 -2  0 -1  1  1  1  0  1  0 -2 -1  0  2 -2  0 -3 -1  0 -2
  [77] -1  1 -2 -1 -1  0 -3  0 -1  2 -1 -3  0 -1  0  0 -4 -1 -1  1 -3 -8 -2 -1 -6 -5 -5  1 -2 -2 -3 -3 -1  0  1 -1 -3  0
 [115] -1  0 -5 -2  2 -2  3  2  3  0  2  0  2 -2 -2 -2 -2  0  0 -1 -1 -2  0  0  0  0 -1  0  4  0 -1  2  1  2  0 -1  0 -1
 [153]  0  0  0 -1  0 -1  0 -1  0  1  1 -3 -1  0  0 -1  0 -2  0  3  0  0 -1  2  1  0  1 -1 -1 -1  1  0 -1  2  0 -1  1  1
 [191]  3  1  0  0 -2  0 -2  2  1  0  0  0  0  0  0 -1  1  1  1  1  1  0 -1  0  0  0  0 -1 -1  1 -1 -1  0  0  0 -1  0  2
 [229]  0 -2 -3  0  0 -1 -1  0  0  2  1  1  0  3  1  0 -4  0  0  0  1  0  0  0  0  0  1  2  1  2 -5  0  0 -1 -1 -3 -3 -6
 [267]  1 -2  0  0  0  2  1 -1  0  0  0  2  1  0  1  2 -2  1  1  0 -1  0  0  0  2  1 -1  0  1  0  1  0  0 -2 -1  0 -1  1
 [305] -1 -2  0  1  1 -1 -1  0 -1 -1  2  0  4 -2  5 -1 -1 -1  0  0  1  1 -1  1 -1  0  0  1  1  1  2  0  0 -3  1 -1 -3  2
 [343] -2  1 -2 -3 -3 -2 -4 -6  0  1 -3  0  2  1 -1  1 -3  0 -2 -1 -1 -1 -1 -2  1  2  7 -1 -4  1 -1 -1 -1 -1  1 -1 -1  0
 [381] -3  2  0  2  0  1  1  0 -1  1  1  0 -2  1  0  0  1  0  0  0  0  0  0  0  2 -1 -1  0  1  1 -2  0  0 -2 -1 -3 -4 -1  0
 [419]  1  1  0  1  0  0  0 -1  0  0  1  0  0 -1 -2  0  1  0  2  3  0  0  0  2 -1  1 -1 -1 -1 -2  1  1  4 -2  1 -4  0  2
 [457]  1  0  0 -1  0  0 -3 -1  0  0  0  0  0  0  0 -2 -2 -1 -1  1  0 -1  0 -2  0 -1 -2 -2 -2 -2 -2 -1 -2 -1  0  0 -1  3
 [495]  1  1  2 -1 -1 -1 -1  2 -1  0 -1  1  0  2 -1  0  0  0 -1  1 -1  2 -1 -1 -1  2  1  0 -1 -1 -5  0  0  0  0  0  0  1
 [533] -2  0 -1  0 -1 -2  0  2 -7  0 -1 -5  2 -1  0  0  4  1  3  1 -3  0 -1  0  1  1  2 -1  0  1  0  3  0  4 -1 -5  0  1
 [571]  1  0  2  0  1  0 -1  0 -3 -1  3  0 -1  2 -1  0 -2 -4  0 -7 -3 -3  1  0 -1  0  0  0  1  0  5 -3  1 -1  0  0 -1  1
 [609]  0 -1  1  0  0  3  0  0 -3 -3  0 -3  0 -2 -1 -4  0 -2  0  0 -1  1  1 -1 -2  0 -2  0 -3  1 -1  0  0  0 -1  1 -6 -1
 [647]  2 -1 -1  0  2  1 -3 -2 -1  0  0  3  1  2  0  0  0  1 -3  0 -1  1  1  0  1  0  0 -1 -4 -3 -1  0  0  2  1  1 -1  0
 [685]  0  0 -1 -1  0  0  2  4  0  1  2  0  0  0  0  1  0  0  1  0  0 -1  0 -1 -3 -4  0 -2  0  0  0  2  2  0  2  1  1 -1
 [723]  2 -1  0 -1  1 -1 -1  1  1  0 -3  1  1  1  0  0 -3 -2 -2 -1 -2  0  0  0  0  0  2  2 -2 -1  2  0  2  0  0  0  0  0
 [761]  1  0 -5 -1  0  0  1 -2  0  2  0  0  2  0  0 -1  0  1  2  0 -1 -4 -4 -2  1 -2  0  0  0  0  1  0  1  0 -1  2  2 -2
 [799]  1  1 -1  1  0  0  0  8 -4 -1  0  0  0 -2  0  2  0 -1  0  0  0  1 -3  2  1  0  0  1  0 -1  0 -1  0 -2 -1  1  1  1
 [837]  4 -3  1  1  0  2  1  1  4  3  3  1  0 -1  1 -1  1  0 -1  0  2  0  4  1  6  0  0  0  0  0 -1  0 -3  1  0  0  0  0
 [875] -3 -3 -3  1  0  0 -1 -5 -1  0  0  2 -1  1  0 -2  0 -1  3 -2  1  0  0 -1  0  1 -1  0  2 -1  1  0 -1  1 -1 -2 -1 -2
 [913]  1  0  0  1  2  0  0  1  0  0  1  1  1  0 -1 -3  1 -1  0 -1  1  5  1  0  2  0 -2  1  0  0  0  0  0  0  0  0 -3  0
 [951]  0  2  0 -1 -1  0  0  0 -1  1  1  1  1 -2 -1 -1 -1  0 -1  0  0 -1  0  1  0  0  0  0 -3  0  0  0  0  1  1  0  1  0
 [989]  0  1  2 -1 -1  1 -2 -1  1  0  0 -1
 [ reached getOption("max.print") -- omitted 1321 entries ]
```

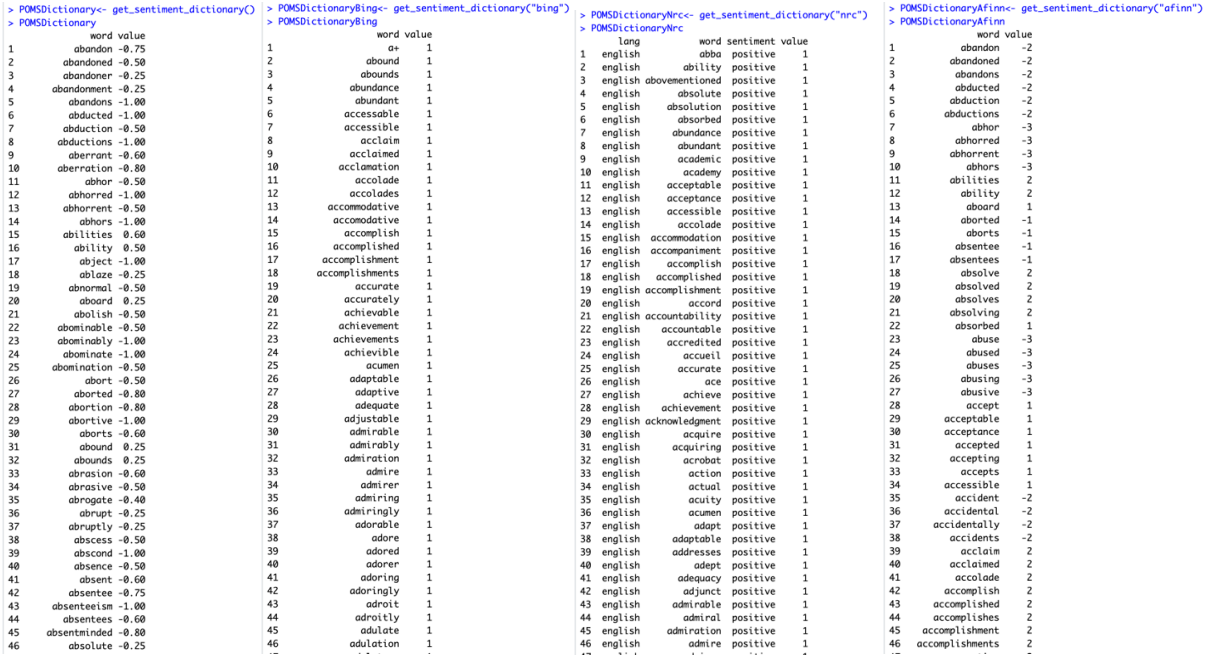
For “afinn” we get the following results.

```
> POMSAfinn<- get_sentiment(POMS, "afinn")
> POMSAfinn
   [1]   0   0   0  -6  -8   2  -2   0  -3   1   2  -1   0   0   7   2  -1   5   1   3   6   0   1  -2  -8   4  -1   1
  [29]   0   0   4   4   3  -1   2   3  -1  -1   0   0   0   4   0   0  -1  -2   5   3  -2  -1  -2   1  -4  -2  -5   1
  [57]   4  -2   0   0  -2   3   0  -1   0   0  -2  -1  -2   7   0  -1  -3   3   1   2  -1   3  -5  -2   0   2  -2   0
  [85]   0   1  -2  -1  -1   0  -2   0  -9  -2  -2  -2  -5 -12  -2   0  -3 -11 -10  -1   1   0  -1  -3  -2   0   0   0
 [113]  -4  -2  -1   0  -3  -5   3   0   3   1   7   4   0   0   2   1   0  -1  -3   0  -1   0   0   5   0   1   0  -2
 [141]  -3   0   1   1   2   1   3   5   0  -1   0   0   0   0  -1   0   0   0   0  -5   0   0   0  -5  -2  -2   3   0
 [169]   1   0   0   2   0   0   1   2   0   0  -3   1  -1  -1   5  -1  -2   0   0   0   0   2   7   0   1  -1   1   0
 [197]   0   5   0  -2   0   0   0   1   0   0   2   6   0   3   2   0   0  -2   0   4   0  -1  -3   2   0   0   0   0
 [225]   0  -2   0   0  -2  -3 -12  -3  -2  -3  -2   3   0   6   1  -1   0   3   4   4  -3   0   0   0   1   1   1   0
 [253]   5   0   6   8   5   2  -9   1   1  -1   2   0   3  -3   5  -3   1   2   0   8   1   1  -2   0   0   2  -1   0
 [281]   1   3   0   0   2  -2  -2   0   2   0   7   3  -3   0   2  -2   0   0   1   0  -1   1  -2   5   0  -2   0   3
 [309]   1   1   1   0   0  -1   0  -1   6  -1   7   2   0  -2   3  -5   0   1  -1   3   0   1   0   3   0   2   2   0
 [337]   0  -2   3   1   2   4  -4   3  -1  -5  -2  -5  -5  -4   0   5   0   0   3   3  -2   2  -3   0   1   0   1   2
 [365]   0  -3   0   0  20  -1   0   2   4  -2  -1  -4   0  -2  -5   1  -5   6   0   0   1   2   5   3  -2   2   3  -3
 [393]  -3   2   0   0   1  -1   0   0  -1   0  -1   4   1   0  -3  -2   3   1   0   3  -8   1  -7  -6  -3   0   2   3
 [421]   0   2   0  -2  -1  -4   0   0   2   0   0  -3  -1   0  -1   2   4   9   0   0  -1   5   2   3   0   0   0  -3
 [449]  -1   2  -1  -3  -2  -7  -3   2   3   0  -1  -5   0  -2   0  -9   3   0   0   0   1  -1   0  -3  -1  -2   2   0
 [477]  -3   0  -1   0   3   0  -3  -8  -2  -2   1   0  -3  -2   2   0  -2   5   2   8   1  -1  -5  -1  -1   0  -1   2
 [505]   0  -1   0   3  -2   0   0  -2  -2   2   1   4   1  -2  -1  -2   2   1   0   0  -1   1   0   0   3   0   0   3
 [533]  -2  -2  -2  -1  -1  -4   0   1  -6   0  -7  -7   4  -3   0   0   8   2   5   5  -9  -2  -1   1   2   7   8   0
 [561]  -2   9   1   8   2   7   0 -13  -2   0   3  -4   6  -2   1   0   0   2  -4  -2   7   0  -2   6  -2   0  -5  -4
 [589]   0 -12 -10  -2   1   0  -2   0   0   0   0   1   3  -8   0  -3  -3   0  -3   0   0  -3   3   0   0   2   1  -4
 [617]  -4   0   3   0   3  -2   2  -4  -1  -2   2  -1  -2  -1   1   0  -5   0  -1   0  -2   1  -2   0   0   0   0  -1
 [645] -10   1   5  -3  -3   0   8   2  -5  -6  -3   2   2   6   3   2   0   0   0   1  -4   0   0   3   0   0   0   3
 [673]   1  -2  -7  -4  -3   0   0   6   3   0  -3   0   1   1  -3   0   2   0   2   6   0   2   5   0   0   2  -5  -2
 [701]  -1   0   2   0   0  -4   0  -1  -4  -3   0  -3  -2   0   2   7   4   0   8   3   3  -2   5  -3   2  -3   2   1
 [729]   0   3  -2   0   0   3   5   0   0  -1  -6   0  -1   1  -2   0  -1   3   0   3   5   5   0  -3   1   0   7   0
 [757]   1   2   0   0   3  -1  -5  -3  -2  -1   6  -6   4   6  -1   0   0   0   0   0   2   2   1  -1  -3  -5  -4   0
 [785]   5  -5   1   0   0   0   5   2   1   0  -2   1   4  -1   1   0   0  -1   0  -1   4  17  -7   0   0   0   0  -1
 [813]   0  -2  -1   3  -1  -1   3   3  -4   3  -4  -3   2   3  -1  -4  -2   0   3  -4  -2   4   4   5   7  -4   0   6
 [841]   0   5   2   0   1   5   4   2   2   0   6  -2  -2   0  -1  -2   3   2   6   3  16   1  -1   1   1  -1  -4   0
 [869]  -1   0  -4  -3   0   0  -3  -3  -2   0  -2   2  -8  -2   1   1   9  -3   3   0  -1   0  -3   4  -3   0   0
 [897]   1   0   3  -1  -1  -2   1   0  -1   0  -3   3  -2  -6   1   0   0   0   0   2   4   2   0   3   0   1   0   0
 [925]   3  -1   2   0   3   9  -1   1   3  13   2   0   1   0  -4   2   0   0  -2   1   0   0   0   0   0  -1   0   2
 [953]   0  -5   2   2   0  -1  -2   4   0   1   3  -5  -1   3  -3  -1   1  -1   0   0   0   0   0   1   0   0  -5  -6
 [981]   0   1   1   5   1   0   0   0   3   3   4   5   0   2  -1  -1   0   0   0   0
 [ reached getOption("max.print") -- omitted 1321 entries ]
```

Lastly, for “nrc”, we get the following results.

```
> POMSnrc<- get_sentiment(POMS, "nrc")
> POMSnrc
   [1]  0  1  0 -1 -4  0  3 -1  2  4  0 -1  1  2  2  0 -1  2  1  3  0  0  1 -1 -5  0  2  2  0  1 -1  3 -1 -1  1 -1  0  0
  [39] -1 -2  4  2  1 -1 -2 -1  1  1 -3  0 -2 -2 -1 -1  0  2  1  1  1  0 -1  1  2 -1  0  0  0 -1  0  5  0  0 -1 -2  3 -2
  [77]  0  3 -2 -2  0  0 -4  0 -1  1 -2  1  1 -1  1  1 -2 -1 -1 -2 -3 -6  2  0  2 -3 -3  3 -3 -2  0  0 -1  0  0 -2 -1  0
 [115]  0  1 -3 -3 -1 -2  1  1  3 -2  3  3 -1 -1  0 -2  0  1  1  1 -2  4  1  0  5 -1 -1  3  2 -5  2  1  2  2  0 -2  0  1
 [153]  0  1  1  3  0  0  1 -3  1  2  2  1  0  1  1 -3 -1 -3  1  1  0 -1  0  2  3  0  1  0  0  0  3  5 -1  4  0  1  1  1
 [191]  4  2  0  1  1  1 -4  1 -1  0  1  2 -3  0  0 -1  2  3  1  1  1  1 -1  0 -1  1  0 -1  0 -1  0 -2  1  0  0 -2  0  0
 [229] -1 -2 -4 -2  1  1  1  1  1  3  3  1  0  0  3  0 -4  2  0  1  0  2  2  1  2  2  3  5  2  2  0  0  2 -2  4 -1 -3 -1
 [267]  3  1  0  1  1  3  0 -1  0  1  1  3  1  0  0  1  0  0  1 -1  0 -2  2  0  4 -1  0  0  2 -2  2  0  1 -1 -1  0  1  3
 [305]  0  1  0  2  1  3  1  0  1  0  2 -1  4 -1  5  0 -3 -1  1 -1 -1  0 -2  3  0 -2  1 -2  2  1  0  1 -1 -4 -1  0 -3  1
 [343] -2  0  1 -1 -2  1 -3 -1 -1  1 -2  1 -1  2  0  2 -1  1 -2 -1  2  1  0  0  0  2 10  0 -1  1  0  1 -2 -1  2  0  0 -3
 [381]  2  6  1  0  2  0  0  2  2 -1 -3  1 -1  4  3 -1  1  1  1  0  1 -1  2  3  2  0  1  2  0 -2 -1  1 -1  3 -4  0  2  0
 [419]  3  2  0  2 -1  0  0  0  0  3  1  0  1  0 -2  0  1  1  4  2  0  2  1  2  1  3  1  1  0 -1  2  1  1 -3 -1 -5  2  2
 [457]  1  1  0 -1  1  0 -1 -3  2  0  0  1  2  0 -1  0 -1 -1  1  1  1  0  1  0 -3  0  1 -5 -1 -2  1 -1 -1  0  2  1  0  3
 [495]  1  3  1  0 -3 -1  1  0  2  3 -2 -2  0  3  1 -1  0  0 -1  1  0  2 -1  1 -3  1  1  0  0  1 -2  2  0 -1  0  0  0 -1
 [533] -1  1 -1 -2  0  0  0 -1 -5  0 -2 -2  3  0 -1  0  4  1  2  1 -1  1  2  1  1  3  0  0 -2  1  0  3  0  3  0 -4  1  2
 [571]  3  2  4 -2  0 -1  1  2 -2  0  4  1 -1  1  0  0 -2 -2  0 -8  0 -2  2  0 -1  1  1  0  2  0  2 -2  1 -2  2  0 -1  1
 [609]  1 -1  0  0  1  2  2 -2  3 -2  2 -1  0 -1 -2 -3 -1 -2 -1 -2 -1  3  1 -1 -2  0  2  0  1  2 -1  1  0  0 -1 -1 -5  0
 [647]  4  1  0  0  4  3 -2  0 -1  1  1  4  1  3  0  0  1  0  0  1  1  1  2  1  0  2  0  1 -4 -3 -2  3  1  1  2  1 -1  0
 [685] -1 -1 -1  2  2  0  0  4  0  1  1  0  2  0 -2  0 -1  0  0  0  0  0  0 -1 -2  1  0 -1  0  0  0  3  2 -1  3  1  1  0
 [723]  4  0  2  1  3  0  0  3  1  0  0  2  2  0  1  1 -2 -3  0  2 -1  0  0  1  0 -1  2  3 -1  1  2  2  2  0  1  0  0  3
 [761] -1  1  0 -1 -1 -2 -1 -1  0  3 -2  0  1  0  0 -1  0  1  4  1 -1 -5 -3 -3  1 -2  0 -1  0  0  1  1  2  3 -1 -3  3 -2
 [799] -1  1  1  1  0  1 -1  7 -3  5  0  1  0  0  2  1  2  2  5  0  4  0  0  0  3 -1 -1  0 -1 -3 -4 -3  2  3  1
 [837]  2 -2  1 -1  0 -1  1  1  4  3  1  1  1  3  3  0 -1 -1 -1 -1  0  2  4  2  6  1  1  0  3 -1 -2 -1 -2  0  0 -2  0  1
 [875]  1 -2 -2 -1 -1  0  0 -4  2  1  1  5  1  2  0 -2 -1  0  2 -1  0  1  1  2  0  0  0  0  2  0  1  1 -1  0 -1 -1 -1 -2
 [913]  1  0  0  1  0  1  2  1  0  0  1  1  1  2  1  0  1  1 -2  1  5  3  0 -1  1 -1  0 -1  0  0  0  0  0  0 -1  1  0
 [951]  0  1  0 -2  1  1  0 -1 -1  1  0  1  0 -4 -3 -2  0  0  0  0  2  0  0  1  0  0  0  0  1  1  1  0  1  4  0  0  0  0
 [989]  1 -1  0  0 -1  1 -1  0  1  2 -1  1
 [ reached getOption("max.print") -- omitted 1321 entries ]
```

Let us look at the sentiment dictionaries for these four methods.



```
> POMSDictionary<- get_sentiment_dictionary()
> POMSDictionary
             word value
1         abandon -0.75
2       abandoned -0.50
3       abandoner -0.25
4      abandonment -0.25
5        abandons -1.00
6         abducted -1.00
7       abduction -0.50
8       abductions -1.00
9        aberrant -0.60
10      aberration -0.80
11          abhor -0.50
12       abhorred -1.00
13      abhorrent -0.50
14         abhors -1.00
15      abilities  0.60
16        ability  0.50
17         abject -1.00
18         ablaze -0.25
19       abnormal -0.50
20         aboard  0.25
21        abolish -0.50
22     abominable -0.50
23     abominably -1.00
24      abominate -1.00
25     abomination -0.50
26          abort -0.50
27        aborted -0.80
28       abortion -0.80
29       abortive -1.00
30         aborts -0.60
31         abound  0.25
32        abounds  0.25
33        abrasion -0.60
34       abrasive -0.50
35       abrogate -0.40
36         abrupt -0.25
37       abruptly -0.25
38        abscess -0.50
39       abscond -1.00
40        absence -0.50
41         absent -0.60
42        absentee -0.75
43   absenteeism -1.00
44      absentees -0.60
45   absentminded -0.80
46       absolute -0.25
```

```
> POMSDictionaryBing<- get_sentiment_dictionary("bing")
> POMSDictionaryBing
             word value
1              a+     1
2          abound     1
3         abounds     1
4        abundance     1
5        abundant     1
6      accessable     1
7      accessible     1
8         acclaim     1
9       acclaimed     1
10     acclamation     1
11       accolade     1
12      accolades     1
13   accommodative     1
14    accomodative     1
15      accomplish     1
16    accomplished     1
17  accomplishment     1
18 accomplishments     1
19        accurate     1
20       accurately     1
21       achievable     1
22      achievement     1
23     achievements     1
24       achievible     1
25           acumen     1
26       adaptable     1
27        adaptive     1
28        adequate     1
29      adjustable     1
30       admirable     1
31       admirably     1
32      admiration     1
33          admire     1
34         admirer     1
35        admiring     1
36      admiringly     1
37         adorable     1
38            adore     1
39           adored     1
40          adorer     1
41         adoring     1
42       adoringly     1
43           adroit     1
44         adroitly     1
45          adulate     1
46         adulation     1
```

```
> POMSDictionaryNrc<- get_sentiment_dictionary("nrc")
> POMSDictionaryNrc
      lang           word sentiment value
1  english           abba  positive     1
2  english        ability  positive     1
3  english abovementioned  positive     1
4  english       absolute  positive     1
5  english     absolution  positive     1
6  english       absorbed  positive     1
7  english      abundance  positive     1
8  english       abundant  positive     1
9  english       academic  positive     1
10 english        academy  positive     1
11 english     acceptable  positive     1
12 english     acceptance  positive     1
13 english     accessible  positive     1
14 english       accolade  positive     1
15 english  accommodation  positive     1
16 english  accompaniment  positive     1
17 english     accomplish  positive     1
18 english   accomplished  positive     1
19 english accomplishment  positive     1
20 english         accord  positive     1
21 english accountability  positive     1
22 english    accountable  positive     1
23 english     accredited  positive     1
24 english        accueil  positive     1
25 english       accurate  positive     1
26 english            ace  positive     1
27 english        achieve  positive     1
28 english    achievement  positive     1
29 english acknowledgment  positive     1
30 english        acquire  positive     1
31 english      acquiring  positive     1
32 english       acrobat  positive     1
33 english        action  positive     1
34 english        actual  positive     1
35 english         acuity  positive     1
36 english         acumen  positive     1
37 english          adapt  positive     1
38 english      adaptable  positive     1
39 english      addresses  positive     1
40 english          adept  positive     1
41 english      adequacy  positive     1
42 english        adjunct  positive     1
43 english     admirable  positive     1
44 english        admiral  positive     1
45 english     admiration  positive     1
46 english          admire  positive     1
```

```
> POMSDictionaryAfinn<- get_sentiment_dictionary("afinn")
> POMSDictionaryAfinn
             word value
1         abandon    -2
2       abandoned    -2
3        abandons    -2
4         abducted    -2
5       abduction    -2
6       abductions    -2
7           abhor    -3
8        abhorred    -3
9       abhorrent    -3
10         abhors    -3
11      abilities     2
12        ability     2
13         aboard     1
14        aborted    -1
15         aborts    -1
16       absentee    -1
17      absentees    -1
18         absolve     2
19       absolved     2
20       absolves     2
21       absolving     2
22        absorbed     1
23           abuse    -3
24         abused    -3
25         abuses    -3
26        abusing    -3
27         abusive    -3
28          accept     1
29      acceptable     1
30      acceptance     1
31        accepted     1
32       accepting     1
33         accepts     1
34      accessible     1
35        accident    -2
36      accidental    -2
37     accidentally    -2
38       accidents    -2
39         acclaim     2
40       acclaimed     2
41        accolade     2
42       accomplish     2
43     accomplished     2
44     accomplishes     2
45   accomplishment     2
46  accomplishments     2
```

We can notice that each of these methods have different criteria and rules that assign sentiment to words, and their dictionaries have different lengths.

Let us sum the values of the sentiment vectors in order to get a measure of the overall emotional valence in the text:

```
> POMSSum<- sum(POMSSentiment)
> POMSSum
[1] 63.3
> POMSBingSum<- sum(POMSBing)
> POMSBingSum
[1] -393
> POMSAfinnSum<- sum(POMSAfinn)
> POMSAfinnSum
[1] 121
> POMSNrcSum<- sum(POMSnrc)
> POMSNrcSum
[1] 556
```

We can observe that except Bing, the other methods say that the content of the text is positive. Bing on the other hand, gives a very strong negative number for the sentiment. To do sentiment analysis, it is important to understand the basis of the sentiment dictionary we use. Since A Princess of Mars is a book, it would be best to use syuzhet as it was designed for analyzing literary texts.

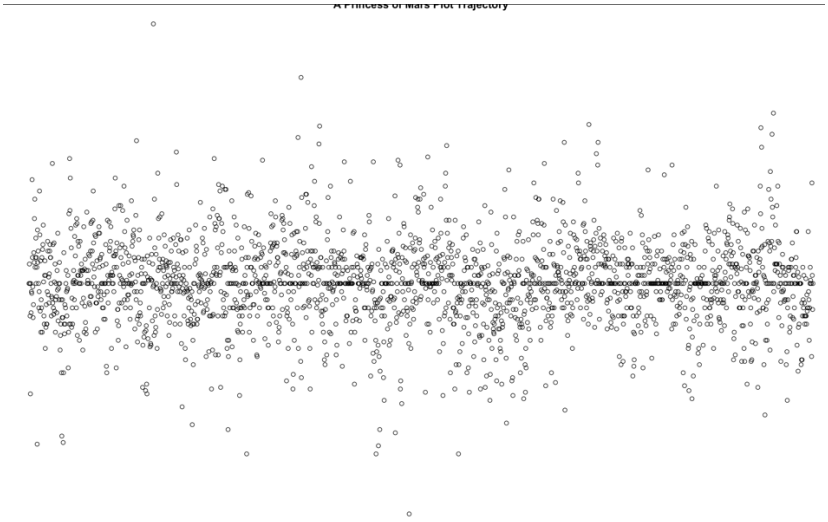
We can then examine the summary() results of the sentiments generated from these dictionaries to understand the distribution of sentiment.

```
> summary(POMSSentiment)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-7.10000 -0.70000  0.00000  0.02727  0.75000  8.00000
> summary(POMSBing)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-9.0000 -1.0000  0.0000 -0.1693  1.0000  8.0000
> summary(POMSAfinn)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-19.00000  -1.00000   0.00000   0.05213   2.00000  20.00000
> summary(POMSnrc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.0000 -1.0000  0.0000  0.2396  1.0000 10.0000
```

The mean sentiment in general, appears to be more neutral. The first half seems to be more negative than the second half. However, to further understand how the sentiment of text changes from the beginning of the text till the end, let us plot the values.

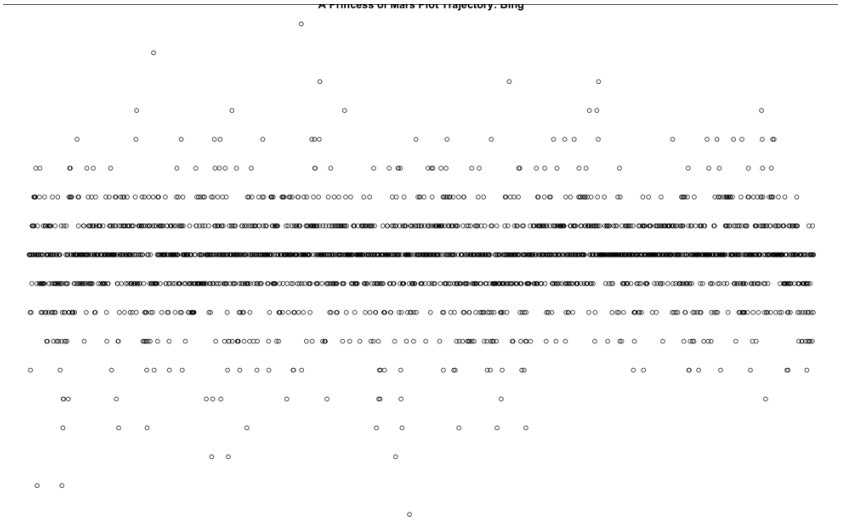
Syuzhet:

```
> plot(POMSSentiment, main="A Princess of Mars Plot Trajectory", xlab="Narrative", ylab="Emotional Valence")
```



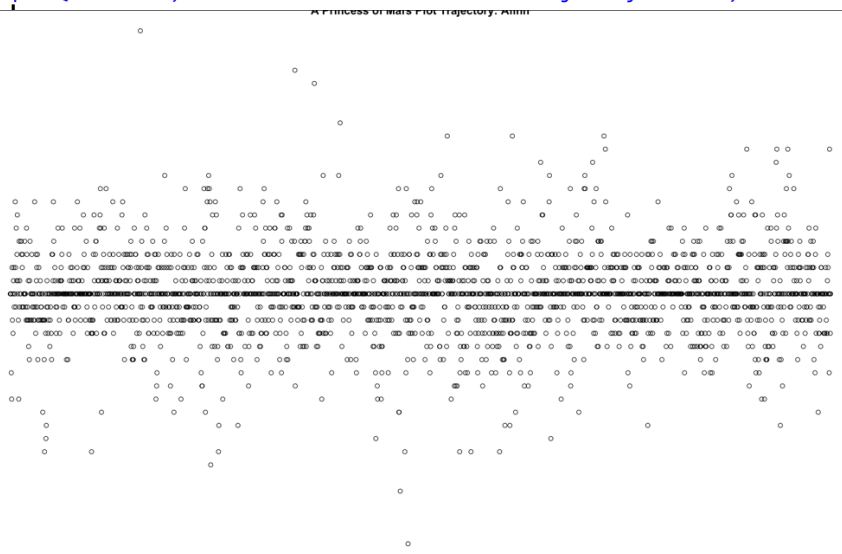
Bing:

```
> plot(POMSBing, main="A Princess of Mars Plot Trajectory: Bing", xlab="Narrative", ylab="Emotional Valence")
```



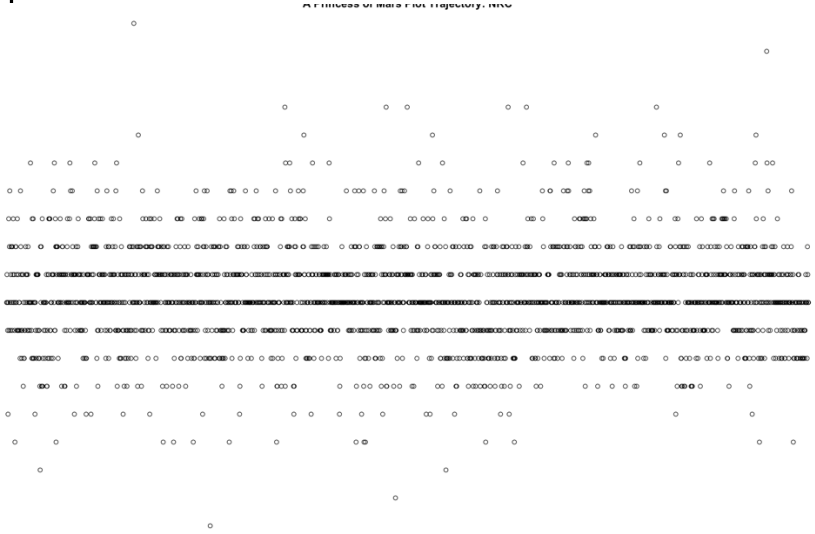
Afinn:

```
> plot(POMSAfinn, main="A Princess of Mars Plot Trajectory: Afinn", xlab="Narrative", ylab="Emotional Valence")
```



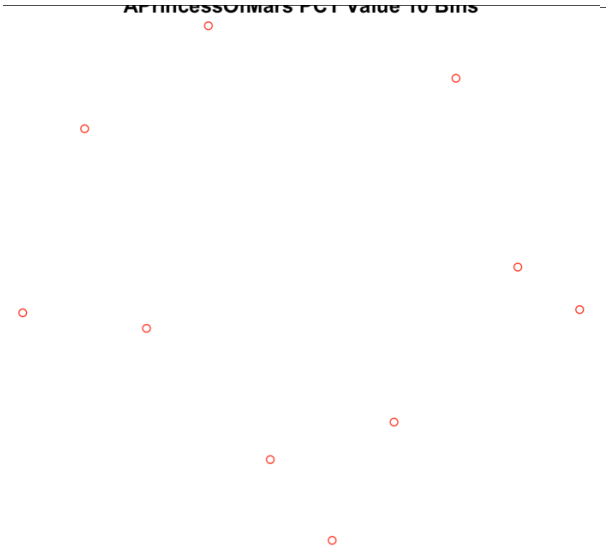
NRC:

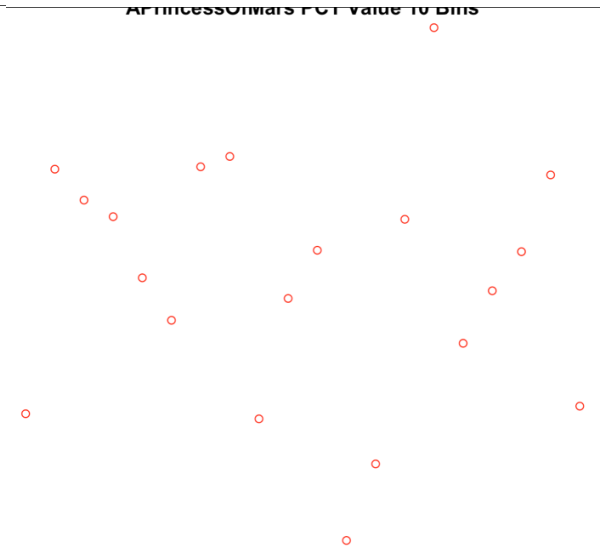
```
> plot(POMSnrc, main="A Princess of Mars Plot Trajectory: NRC", xlab="Narrative", ylab="Emotional Valence")
```

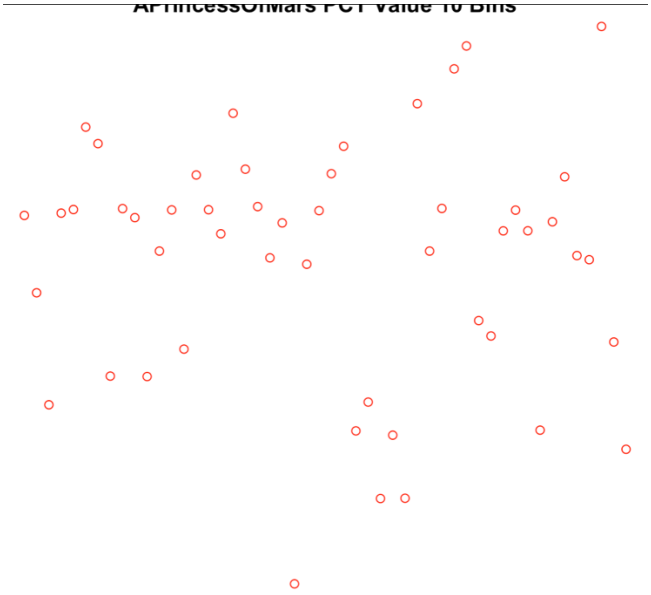


Let us now use the get_percentage_values() function from the syuzhet package to compare the shape of one trajectory to another. We have plotted values for 10, 20 and 50 bins respectively.

```
> plot(POMSnrc, main="A Princess of Mars Plot Trajectory: NRC", xlab="Narrative", ylab="Emotional Valence")
>
> POMSSentimentPctValue <- get_percentage_values(POMSSentiment, bins=10)
> structure(POMSSentimentPctValue)
          1           2           3           4           5           6           7           8           9
 0.006437768  0.156681034 -0.006250000  0.240732759 -0.113362069 -0.179310345 -0.082758621  0.197844828  0.043750000
         10
 0.009051724
> plot(POMSSentimentPctValue, main="APrincessOfMars PCT Value 10 Bins", xlab="Narrative", ylab = "emotional valence", col="re
d")
```







Through these plots, we understand that the sentiment shifts a lot throughout the text, from being very high in the beginning, to low in the middle, and high again towards the end. This can imply volatile and more emotional content in the book. The large number of sentences, prominent words in the word cloud such as martian, creature, incubator, spear, plateau, city, warriors etc. hint towards a science-fiction themed text with violent acts being performed throughout, further hinted by the volatile sentiments. However, further analysis may be required to understand the text on a deeper level.

6. Discussion

This project helped us gain practical hands-on experience with Text Analytics. Following the rubric, we had a path to follow for this project, yet the challenges along the way, such as dealing with multiple documents, or accommodating large text, reducing sparsity etc. are all concepts we learned while trying to solve a problem. For the same reason, we also understood why Data Science is an empirical science. We learned about data preprocessing, and how cleaning the data before perfoming the analysis is of such a high importance, because stopwords, punctuation, quote marks etc. are all things that do not contribute to the analysis and thus, need to remove as keeping them could skew the analysis we are trying to perform. We learned various packages for Text Analytics and the interesting capabilities they support, including creating wordclouds, retrieving information about the text, assigning sentiments etc. Through all these bits and pieces of analysis, in the end we are able to gain a somewhat rough image of the text we are looking at.

The Text Analytics project was a great learning experience and as the third project, accumulated up the knowledge we gained from the previous projects. It helped solidify approaches, thought process and troubleshooting for analytics. It also proved yet again the convenience R brings to the table when it comes to performing high quality comprehensive data analytics.

To summarize, after the completion of this project, we are confident in working with R, performing text analytics on large datasets, cleaning data, extracting useful information, generating plots, wordclouds, understanding sentiments, applying different functions and measures/metrics on the documents that lets us draw different inferences about the problem domain i.e. the text, along with simplifying text for better analysis.